

Nonparametric statistical inference for functional brain information mapping

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr. rer. nat.)

im Fachgebiet
Informatik

vorgelegt
von Dipl.-Phys. Johannes Stelzer

geboren am 21.04.1983 in Schwäbisch Hall

Die Annahme der Dissertation wurde empfohlen von

1. Prof. Dr. Martin Bogdan (Universität Leipzig)
2. Prof. Dr. Nikolaus Kriegeskorte (University of Cambridge, GB)

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 16.04.2014 mit dem Gesamtprädikat magna cum laude

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder erbrachten Dienstleistungen als solche gekennzeichnet.

Kopenhagen, den 21. Mai 2014

.....
(Ort, Datum)


.....
(Unterschrift)

Bibliographische Daten

Stelzer, Johannes

Nonparametric statistical inference for functional brain information mapping

Juni 2013

Max-Planck-Institut für Kognitions- und Neurowissenschaften Leipzig, Dissertation

154 Seiten, 132 Literaturangaben, 50 Abbildungen, 4 Tabellen

Wissenschaftlicher Werdegang des Verfassers

- **2009-2013:** Doktorarbeit am Max-Planck-Institute for Human Cognitive and Brain Sciences (Neurophysics Department / Music Cognition and Action Group), Leipzig, Deutschland
- **2005-2009:** Studium der Physik (Diplom), Biophysik und Philosophie, Universität Leipzig, Deutschland
- **2003-2005:** Studium der Physik (Vordiplom), Mathematik und Philosophie, Universität Hamburg, Deutschland

Contents

Acknowledgments	xi
List of Figures	xiii
List of Tables	xiv
List of Symbols	xvi
I Introduction and background	
1 Introduction	3
2 Magnetic Resonance Imaging	7
2.1 Nuclear Magnetic Resonance	7
2.2 Perturbation and relaxation	9
2.3 Spin-lattice relaxation	9
2.4 Spin-spin relaxation	10
2.5 Apparent spin-spin relaxation	10
2.6 Detection of the MR signal	10
2.7 Position encoding	11
2.8 T_1 and T_2 weighted images	12
3 Functional MRI	13
3.1 Blood as contrast agent for MRI	13
3.2 Biophysics of the BOLD signal	14
3.3 Functional MRI recordings	15
3.3.1 Correlational structure of the data	15
4 State of the art brain analysis	17
4.1 Human brain mapping	17
4.2 Overview of fMRI analysis methods	18
4.3 The general linear model	19
4.3.1 Mathematical formulation of the GLM	19
4.3.2 Criticism of GLM methods	21
4.4 Multivariate pattern analysis	21
4.4.1 MVPA in neuroimaging	21

4.4.2	Classifiers	22
4.4.3	Feature Selection	23
4.4.3.1	Wrapper methods	23
4.4.3.2	Region of interest analysis	25
4.4.3.3	Dimensionality reduction of the feature space	26
4.4.3.4	Searchlight approach	26
4.4.4	Cross-validation	26
4.4.5	Pattern analysis methods beyond classification	27
4.4.5.1	Regression	27
4.4.5.2	Encoding	28
4.4.6	Criticism of MVPA methods	28
4.4.6.1	Decoding and regression	28
4.4.6.2	Encoding	28
4.5	Activation vs. Information mapping	29
4.5.1	Searchlight decoding	30
4.5.2	Feature weight mapping	30
5	Statistical inference	31
5.1	Hypotheses testing	31
5.2	Parametrical statistical inference	34
5.2.1	Z-test	34
5.2.2	T-test	35
5.2.3	Binomial models	36
5.2.4	Tests for normality	37
5.3	Nonparametric statistical inference	37
5.3.1	Permutation tests	37
5.3.2	Bootstrapping methods	39
5.4	The multiple comparisons problem	40
5.4.1	Bonferroni correction	43
5.4.2	False discovery rate	43
5.4.3	Random field methods	44
5.4.4	Non-parametric cluster statistics	45
II	Methods	
6	fMRI data sets	49
6.1	3T tapping synchronization experiment	49
6.1.1	Experimental design	49
6.1.2	Data acquisition	50
6.1.3	Data preprocessing	50
6.2	7T finger tapping and imagination	51
6.2.1	Experimental design	51
6.2.2	Data acquisition	51
6.2.3	Data preprocessing	53
7	Synthetic data sets (simulations)	55

7.1	Single Subject Simulations	55
7.1.1	Single subject geometric simulation	55
7.1.2	Single subject null simulation	56
7.2	Group Simulations	56
7.2.1	Group simulation 5 cubes	56
7.2.2	Group null simulation	57
7.3	General simulations	58
7.3.1	Cross-validation influence simulation	58
7.3.2	Simulation undersampling the permutation space	58
8	Preprocessing of the fMRI data	61
8.1	Motion Correction	61
8.2	Temporal filtering	61
8.3	Normalization to standard brain space	63
8.4	Temporal bundling of scans	63
9	Multivariate Analysis & Statistics	65
9.1	Support vector classification	65
9.2	Searchlight decoding (SLD)	68
9.3	Feature weight mapping (FWM)	69
9.4	Permutation testing	70
9.5	Group level Monte-Carlo recombination	70
9.6	Threshold map procedure	71
9.7	Cluster size statistics	71
9.8	Parametric framework for comparison	72
9.9	Processing pipelines	73
9.9.1	SLD on single subject level	73
9.9.2	SLD on the group level	74
9.9.3	FWM on single subject level	75
9.9.4	FWM on the group level	76
III	Results	
10	Singe subject results	81
10.1	Single subject geometric simulation	81
10.1.1	Qualitative comparison between the FWM and SLD method	81
10.1.2	Influence of geometry	86
10.2	Single subject null simulation	87
10.2.1	Influence of underlying image smoothness	87
10.3	3T tapping synchronization experiment	88
10.4	7T finger tapping and imagination	93
11	Group analysis results	99
11.1	Group simulation 5cubes	99
11.1.1	Nonparametric vs parametric	99
11.1.1.1	Searchlight decoding	99

11.1.1.2	Feature weight mapping	105
11.1.2	Searchlight decoding vs feature weight mapping	107
11.2	Group null simulation	113
11.2.1	Searchlight decoding	113
11.2.2	Feature weight mapping	113
11.3	3T tapping synchronization experiment	115
11.3.1	Searchlight decoding	115
11.3.2	Feature weight mapping	115
11.3.3	Comparison of SLD vs FWM	117
12	General results	121
12.1	Cross-validation influence simulation	121
12.2	Simulation undersampling the permutation space	121
IV	Discussion	
13	Statistics in fMRI	127
13.1	Pitfalls of parametric statistics in classification-based fMRI	127
13.1.1	T-based statistics	127
13.1.2	Binomial models	128
13.2	Characteristics of the nonparametric framework for classification-based fMRI	130
13.2.1	Preservation of spatial structure in chance maps	130
13.2.2	Threshold map procedures	131
13.2.3	Cluster statistics	132
13.2.4	Dependency on the voxel-wise threshold	133
13.3	Comparison between nonparametric and parametric statistics in classification-based fMRI	134
13.3.1	Sensitivity	134
13.3.2	Precision	135
13.3.3	Credibility	136
13.3.4	Conclusion on the quality of nonparametric and parametric tests	136
14	Information mapping methods	139
14.1	Searchlight decoding	139
14.2	Feature weight mapping	142
14.3	Conclusion on information mapping techniques	144
15	Single subject or group studies	145
15.1	Motivation for group level inference	145
15.2	Motivation for the analysis on the single-subject level	146
15.3	Conclusion on level of inference	147
V	Conclusion	
16	Summary	151
16.1	Limitations	152

“The brain, is the most complex thing we have yet discovered in our universe.”

James D. Watson

The techniques introduced in this work aim to assist in further unraveling of the mysteries¹ of the most complex piece of matter that we know, the human brain. This thesis would not have been possible without the help of many individuals, hence I want to take the opportunity here to say thank you. In particular, I would like to thank Professor Robert Turner for his outstanding scientific guidance while giving me freedom to pursue projects of my own choice. My thanks go also to Gabriele Lohmann, Yi Chen and Tilo Buschmann for sharing countless discussions; I especially enjoyed our constant exchange of ideas over the last years. My best thanks go to Mike Hove for being a great office mate, introducing me to shamanism (in particular from a neuroscientific point of view) and sharing his fMRI data. Also thanks to Robert Trampel for sharing his stunning high-resolution data sets and to Masami Ishihara and Peter Keller for introducing me into the world of music psychology. Another big thank you goes to David Smith who corrected my English in this thesis and to Jörn-Henrik Jacobsen for putting together a interesting study on musical memory jointly. Furthermore I would like to thank all here not mentioned staff in the Max-Planck-Institute for being so open and helpful. Thanks to Professor Bogdan from the computer science department of the University of Leipzig for his time and efforts concerning this thesis. To all my friends here in Leipzig and elsewhere in Germany and the world, thank you for your patience and understanding! A very special thanks goes to my family, in particular my father Erwin and mother Madeleine, who enabled us with the greatest education and freedom, my brother Fabian for reminding me of opportunities outside of the field of science and my sister Larissa for reminding me of the importance of music in life. Un muy especial agradecimiento a Virginilla por estar ahí por mí! Furthermore, our almighty saviour, the Flying Spaghetti Monster, shall not remain unmentioned here. Lastly thanks to you, dear reader, for looking into my thesis!

¹With all modesty and in particular considering the sheer complexity of the human brain, it should be noted that in the current state of neuroscience and with the tools currently available (including those introduced here), *most* mysteries of the human brain will remain undiscovered.

List of Figures

3.1	Comparison of a 3 Tesla and a 7 Tesla BOLD fMRI scan within the same subject	15
3.2	Exemplary BOLD time series of one single voxel	16
4.1	Experimental rationale for brain mapping studies	18
4.2	Exemplary BOLD time series with GLM fit	20
4.3	Comparison between linear and non-linear classifiers	24
5.1	Trade-off between sensitivity and precision of statistical testing procedures	33
5.2	Z-distribution with the parameters $\mu_0 = 0$ and $\sigma = 1$	35
5.3	Histogram of exemplary data of health values <i>under permutation</i>	38
5.4	Approximation of exemplary data of health values <i>using the bootstrap method</i>	40
5.5	Visualization of the multiple comparisons problem using three successive dice throws	42
6.1	Illustration showing the difference in resolution of the two fMRI experiments used in this thesis	50
6.2	Brain coverage of the ultra-high resolution fMRI data set	52
7.1	Information spread of the single-subject geometric simulation	56
7.2	Information spread for the group simulation 5 cubes	57
8.1	Effects of subject head motion depends on voxel size	62
9.1	Schematic overview of the nonparametric framework on the single-subject level	66
9.2	Schematic overview of the nonparametric framework on the group-level	67
10.1	Overview of the single-subject geometric simulation results	82
10.2	Threshold maps for the single-subject geometric simulation	82
10.3	Cluster size histogram for the single-subject geometric simulation analyzed by the SLD method	83
10.4	Cluster size histogram for the single-subject geometric simulation analyzed by the FWM method	84
10.5	Precision/recall curves for the three different levels of information distribution	85
10.6	Impact of intrinsic smoothness on cluster-size histograms	88
10.7	Comparison between the SLD and FWM method on the single-subject level for the 3T tapping synchronization experiment	89

10.8	Threshold maps for the 3T tapping synchronization experiment on the single-subject level	90
10.9	Cluster size histograms for the 3T tapping synchronization experiment using the SLD method on a single-subject level	91
10.10	Cluster size histograms for the 3T tapping synchronization experiment using the FWM method on single-subject level	92
10.11	Slice orientation for the high resolution 7T finger tapping and imagination data set	94
10.12	Single subject results for the 7T finger tapping and imagination data set (low threshold)	95
10.13	Single subject results for the 7T finger tapping and imagination data set (high threshold)	96
10.14	Cluster-size histograms for single-subject fMRI at 7T using SLD	97
10.15	Cluster-size histograms for single-subject fMRI at 7T using FWM	98
11.1	Group simulation 5cubes data set analyzed by the SLD method, comparison parametric vs. nonparametric	100
11.2	Searchlight derived nonparametric cluster size histogram from the group simulation 5cubes data set	101
11.3	Influence of the initial voxel threshold in searchlight decoding (group simulation 5cubes)	102
11.4	Influence of searchlight diameter in the group simulation 5cubes	104
11.5	Group simulation 5cubes data set analyzed by the feature weight mapping method, comparison parametric vs. nonparametric	106
11.6	Feature weight mapping derived nonparametric cluster size histogram from the group simulation 5cubes	107
11.7	Influence of the initial voxel threshold in feature weight mapping (group simulation 5cubes)	108
11.8	Comparison between searchlight decoding and feature weight mapping on the group simulation 5cubes	109
11.9	Influence of voxel-wise threshold on the SLD and FWM method in the group simulation 5cubes	110
11.10	Influence of the number of subjects on the SLD and FWM method in the group simulation 5cubes	111
11.11	Overview of the group-level searchlight decoding results of the 3T tapping synchronization experiment, comparison parametric vs. nonparametric	116
11.12	Searchlight decoding derived nonparametric cluster-size histogram group-level analysis of the 3T tapping synchronization experiment	117
11.13	Overview of the group-level feature weight mapping results of the 3T tapping synchronization experiment, comparison parametric vs. nonparametric	118
11.14	Nonparametric cluster-size histograms of the 3T tapping synchronization experiment using FWM	119
11.15	Direct comparison of the group-level results of the SLD and FWM method analyzing the 3T tapping synchronization experiment	120

12.1	Cross-validation influence simulation showing the impact of different cross-validation schemes	122
12.2	Simulation undersampling the permutation space	123
14.1	Schematic illustration of searchlight induced inflations and distortions	140

List of Tables

5.1	Relations between the truth or falseness of the null hypothesis H_0 and the outcomes of a test	32
10.1	Results of the null simulation on single-subject level	87
11.1	Results of group null simulation for the SLD method (parametric versus non-parametric)	113
11.2	Results of group null simulation for the FWM method (parametric versus non-parametric)	114

List of Symbols

BOLD	Blood Oxygen-Level Dependent
FDR	False Discovery Rate
fMRI	Functional Magnetic Resonance Imaging
FWE	Family Wise Error
FWHM	Full Width at Half Maximum
FWM	Feature Weight Mapping
GLM	General Linear Model
GPU	Graphics Processing Unit
LGN	Lateral Geniculate Nucleus
LOOCV	Leave-One-Out Cross-Validation
MR	Magnetic Resonance
MRI	Magnetic Resonance Imaging
MVPA	Multivariate Pattern Analysis
NMR	Nuclear Magnetic Resonance
p-value	Probability value
PCA	Principal Component Analysis
PET	Positron Emission Tomography
RF	Radio Frequency
RFE	Recursive Feature Elimination
ROI	Region Of Interest
SLD	Searchlight Decoding
SPM	Statistical Parametric Mapping

SVM Support Vector Machine
TR Scanner Repetition Time

Part I

Introduction and background

Chapter 1

Introduction

The human brain is a network consisting of a huge number of interacting agents (neurons), which are able to organize themselves into countless functional subnetworks. The local and global interplay of neurons and networks of neurons ultimately makes available *everything* that constitutes our personal reality - from the memories of the past to the endless stream of thoughts and perceptions.

With the advent of techniques such as functional magnetic resonance imaging (fMRI), it has become possible to observe the activity of the living human brain on a macroscopic scale. Brain scanners based on fMRI have the huge advantage that they are fully noninvasive, since blood was found to be a natural contrast agent linked to neuronal activity. Modern fMRI scanners are capable of sampling the dynamics related to the blood flow and hence neuronal activity of the entire brain at an ever-increasing resolution. Recordings on the scale of few cubic millimeters or smaller are possible by now; the recordings are of the form of three-dimensional pixels, known as voxels.

Researchers affiliated with imaging-based human brain sciences commonly aim to construct a *map* of brain function, i.e. to map the functional organization of the brain. The most widespread ways for doing so are *activity-based univariate* analysis methods: in here, each recorded image element (voxel) is analyzed separately (hence the term univariate) in regards to differences of their activity. For instance, the activity level of a voxel in the visual cortex may be systematically higher if visual stimulation is presented as compared to a control condition with no visual stimulation (hence, the voxel in the visual cortex becomes *activated* here).

Recently, more sophisticated analysis methods for brain imaging data from fMRI have become increasingly popular. Many of these new techniques originated from machine learning methods. The main difference between the traditional univariate activation-based methods and the recently adopted machine learning methods is that not only is the activity from *one* voxel solely considered, but the activity of *many* voxels is analyzed simultaneously. Furthermore, the new methods make it possible to delineate differences in brain function which manifest themselves rather in the form of a complex fingerprint of brain activation as opposed to simple voxel-wise differences of activity. Multivariate machine learning approaches also enable a

mapping of brain function; this group of approaches is commonly referred to as *information mapping methods*.

While the newly introduced information mapping techniques often offer a higher sensitivity and enable more detailed insights into patterns of brain activity, the statistical methods for data analysis applied in research often are of a rather premature nature. However, appropriate statistical frameworks are *absolutely indispensable* in the search of robust principles of functional organization, as the brain is an inherently noisy system. Most commonly, researchers use statistical frameworks for the new machine learning based methods that have been developed for the analysis of univariate activation-based methods. This course of action, however, is problematic as the underlying assumptions of the univariate statistical frameworks are unmet if applied here. For exactly this point, my work offers a novel solution, which is fully adapted for information mapping techniques based on machine learning methods.

The solution introduced in my thesis is based on previous resampling-based statistics for classification methods[1, 2, 3, 4, 5], however is extended to offer a full solution for the characteristics of fMRI data (in particular the multiple comparison problem). The proposed framework is capable of analysis both on the level of single subjects but also for group-level analysis.

I demonstrate the applicability of the statistical framework for two distinct *information mapping* techniques (for a more detailed differentiation between information-mapping techniques and activation based brain images see Section 4.5 on page 29):

1. Volumetric searchlight decoding (SLD)
2. Feature weight mapping (FWM) using matrix decompositions

Both information-mapping methods have in common that they attempt to reveal brain regions which are informative about the stimulus condition. The underlying rationale and implementation of the two information mapping methods, however, is strikingly different: while the SLD method uses local neighborhoods of voxels and yields decoding accuracy maps, the FWM method uses pattern information of the whole brain and establishes feature weight maps.

As both of these information mapping techniques have a different methodology and rationale, the scope of this work is two-fold: Firstly, I investigate the characteristics and differences between the two information mapping approaches (SLD and FWM), both in light of the novel nonparametric statistical framework. This characterization is performed using fMRI data and simulations, both on the single-subject and group level. The second scope of my thesis is to compare the proposed nonparametric framework with other *parametric* frameworks for assessing statistical significance that are commonly used for classification-based fMRI (which originated from univariate analysis methods). The comparison between nonparametric and parametric inference, however, is carried out on the group level only, due to limitations in the applicability of parametric methods for single subject inference.

My work is organized as follows: first I will introduce the physical principles behind

magnetic resonance imaging and the biophysical principles allowing the usage of this technique for *in-vivo* functional brain imaging. This is followed by a state of the art review of functional brain imaging techniques including univariate and different multivariate machine learning techniques. Next I will provide an overview over the statistical inference principles and methods tailored for the statistical analysis of fMRI data.

In the methods section I describe the two fMRI experiments that I use in this thesis and the data generation of the simulations that are employed. Furthermore, I describe the methods and pipelines that are used for the nonparametric framework.

The results section is divided into the single-subject and group level. On both levels, the results of fMRI studies and various simulations are shown. The following discussion is split into three sections; in the first section I discuss the theoretical issues with commonly used parametric statistical frameworks when applied to information mapping methods and the further aspects of the novel nonparametric framework. In the following section of the discussion I compare the two information-mapping methods that are used for my thesis (the SLD and FWM method). Lastly, I finish with a short section on the rationale of group studies versus single subject level studies.

Chapter 2

Magnetic Resonance Imaging

2.1 Nuclear Magnetic Resonance

Magnetic Resonance Imaging (MRI) is a relatively new imaging technology, with the first image¹ acquisition taking place in 1973 [6]. MRI is based on the physical principle of Nuclear Magnetic Resonance (NMR), which is a quantum mechanical phenomenon describing how the magnetic cores of atoms, known as nuclei, absorb and emit energy in the form of photons in the presence of an external magnetic field.

The atomic nuclei consist of two kinds of constituents, known as nucleons: protons (which carry electric elementary charge² of $+1e$) and neutrons (which are electrically neutral). The nucleons are bound together by the nuclear force, which on short length scales massively exceeds the electrostatic repulsion that protons otherwise exert onto each other. In the quantum mechanical description, each nucleon possesses a *spin angular momentum* (short: spin). As a fundamental intrinsic property of all elementary particles and their composites, the spin does not have a direct counterpart in the regime of classical physics. The spin angular momentum sometimes is thought to be *analogous* to the angular momentum known from classical mechanics, such as for instance the angular momentum of the rotation of our planet. However, the spin angular momentum is an *intrinsic* quantum mechanical parameter, which is also carried by point-like particles such as the electron, which due to their lack of spatial extension cannot have classical angular momentum.

The spin of a particle determines the possible quantum states that it can occupy. Furthermore, the property of spin causes the nucleus to have a *magnetic moment* (known as the nuclear magnetic moment). The magnetic moment is a vector quantity, incorporating the size *and* direction of a magnetic dipole of the nucleus. Most crucially, the magnetic moment interacts with an external magnetic field B_0 , which for convenience is set here in the direction of the *z-axis*. According to the description of classical mechanics [7, page 65], it can be shown that the spins start to *precess* around the direction of B_0 at whatever deflection angle θ relative

¹image in the sense of an at least *two-dimensional* representation

² $+1e = 1.60217 \times 10^{-19}C$

to the z -axis which they happen to be aligned with. The frequency of the precession is known as the *Larmor frequency*, which depends linearly on the magnetic field strength:

$$\omega_0 = \gamma B_0 \quad (2.1)$$

where γ is the gyromagnetic ratio, which depends on the *type* of nucleus considered. As stated before, in the classical (not quantum mechanical) description, the precession can take any value for the angle θ . Initially the nuclear magnetic moments from a sample cancel each others out and the net magnetization is zero, as the magnetic moments principally underlie the superposition principle [8]. However, there exists a difference in potential energy of a single magnetic moment in the presence an external field; this energy is at a minimum if the magnetic momentum and the external field are parallel ($\theta = 0$) and at a maximum, if they are antiparallel ($\theta = \pi$). When regarding a macroscopic sample, which typically consists of 10^{20} to 10^{26} spins, over the time scale of seconds (which is very slow compared to the frequency of precession), thermal fluctuations cause the spins to slightly favor the lower energy state (according to Maxwell-Boltzmann statistics), until thermal equilibrium is reached[9]. The discrepancy of the occupation depends on the strength of the external magnetic field and the temperature. Note that the equilibrium is dynamic, i.e. the spins are allowed to change their alignment and therefore energy, as long as the total energy of the system remains constant. Given this favoring of occupation levels between the parallel and antiparallel states, a net magnetization M_z in the direction of the external field B_0 (z -direction) is built up. The magnetization in the x - y plane is cancelled out, as there is no preferred alignment.

As demonstrated in the ground breaking Stern-Gerlach experiment [10], the classical approach does not fully capture the physical reality, since the energy of the system can only take certain, quantized values. Consequently, only a discrete set of possible orientations for the angular momentum exist. In the simplest case of a hydrogen nucleus (which consists of a single proton) in an external magnetic field, there are only two possible energy states. This splitting of an energy level in presence of a magnetic field is known as Zeeman effect.

$$E_{\uparrow} = -\frac{1}{2}\hbar\gamma B_0 \quad (2.2)$$

$$E_{\downarrow} = +\frac{1}{2}\hbar\gamma B_0 \quad (2.3)$$

If a proton of the parallel E_{\uparrow} state absorbs the energy ΔE in form of a photon, which carries the energy difference between the two states $\gamma\hbar B_0$, the proton can switch into the antiparallel E_{\downarrow} state. This energy precisely corresponds to the energy that is carried by a photon of the frequency corresponding to the Larmor frequency of precession:

$$\Delta E = \hbar\gamma B_0 = \hbar\omega_0 \quad (2.4)$$

It should be noted that in the full quantum mechanical description, single spins do not have a deterministic direction and energy level, as they are principally in a *superposition* of all possible states (in case of hydrogen the proton would be in both states simultaneously). On a level of a macroscopic sample, an observable population difference between the two energy states arises[9]. The observable population difference between the two energy states causes the sample to have a net magnetization in z -direction.

2.2 Perturbation and relaxation

The underlying idea behind MRI is that the dynamic equilibrium of a spin system is perturbed, i.e. brought into a disequilibrium. Next, the system falls back into its equilibrium state. This process is called *relaxation* and is the key to the MR signal. The relaxation process is governed by the first two laws of thermodynamics as the system transits into a low-energy and high-entropy state. The perturbation of the system usually is implemented by a second magnetic field B_1 , which rotates at the Larmor frequency in the x - y plane. As the Larmor frequency typically is within the wavelength range of radio signals (for the magnetic field strengths commonly used), the perturbation is commonly denoted as radio frequency (RF) pulse.

As a result of this RF pulse, the spins absorb energy and the magnetization of the system develops a component in the x - y plane and the size of the component in the z -direction (the direction of the B_0 field) is decreased. The angle α between the z -axis and the net magnetization of the system is known as the flip angle. All spins are exposed to the same B_1 radio frequency field, hence their magnetization is in phase and rotates around the z -axis. During application of the RF pulse, the spins rotate in the same phase, i.e. their phases are coherent. The RF pulse is only applied for a brief period of time, after which the system evolves back into its equilibrium state.

It is possible to distinguish two kinds of relaxation processes [7, page 69]: Spin-lattice relaxation and spin-spin relaxation.

2.3 Spin-lattice relaxation

The perturbation of the system caused by the application of the RF pulse implies the *absorption* of energy. As the disequilibrium state is thermodynamically unstable, the spin system returns the energy of the RF pulse into the surrounding tissue. In other words the excess energy that previously had been absorbed is dissipated. More precisely, the energy is redistributed into rotational and vibrational degrees of freedom in the surrounding tissue, in the format of heat. This implies a (very small) rise in temperature of the surrounding materials. Over time, the longitudinal component of the magnetization M_z (in direction of the B_0 field) is fully restored. The recovery of the z -magnetization can be approximated[11, page 54] as

$$M_z(t) = M_0 \left(1 - (1 - \cos\alpha)e^{-t/T_1} \right) \quad (2.5)$$

with the flip angle α and the equilibrium net magnetization M_0 in z -direction. The time constant T_1 found in the denominator of the exponential is defined as spin-lattice relaxation time and depends highly upon the type of material surrounding the spins. As an example, the T_1 time for grey matter in the frontal lobes is around 1200ms, while the T_1 time for cerebrospinal fluid is around 4300ms for $B_0 = 3T$ [11, page 50]. The recovery of the longitudinal magnetization hence is much faster for smaller values of T_1 .

2.4 Spin-spin relaxation

The magnetic fields of spins can interact with each other's, therefore it is possible that the Larmor frequency is temporarily shifted. This causes the loss of phase coherence, which implies the decay of the transverse magnetization in the x - y plane. The temporal decay of the transverse magnetic component can be approximated[7, page 69] as

$$M_{tr}(t) = M_0 \sin \alpha e^{-t/T_2} \quad (2.6)$$

where α is the flip angle and M_0 the net magnetization in z -direction in the equilibrium state. The constant T_2 in the denominator of the exponential is known as the spin-spin relaxation time. The T_2 time constant depends highly on the molecular and chemical environment of the spins which itself depends on the nature of the surrounding tissue. For instance, the T_2 time for grey matter in the frontal lobes is around 88ms, while the T_2 time for cerebrospinal fluid is around 1442ms for $B_0 = 3T$ [11, page 50]. Therefore, the transverse magnetization decays much faster for grey matter than for cerebrospinal fluid. In general terms, the transverse T_2 relaxation is faster than the longitudinal T_1 recovery.

2.5 Apparent spin-spin relaxation

In more realistic circumstances, local variations of the magnetic field strength speed up the loss of phase coherence[12, page 159]. There are two factors for different local fields; on one hand inhomogeneities in the B_0 field, and on the other susceptibility effects, i.e. local variations in the degree of magnetization given an external magnet field. The transversal relaxation time incorporating this effect is known as *apparent* spin-spin relaxation time T_2^* , and is generally lower than the T_2 time given the same material.

2.6 Detection of the MR signal

In summary, a perturbation in the form of a RF pulse causes a non-zero transverse component of the magnetization (i.e. in the x - y plane), which decays according to Equation 2.6. The transverse component induces an electric field in the receiver coil of the MR system - which can be measured and is known as the MR signal. When returning to the equilibrium state, the

transverse magnetization is lost and the MR signal decays. Typically, more than one measurement is applied and the time constant between the perturbations is called the repetition time TR. Note that the longitudinal magnetization does not have to be restored to the full equilibrium magnetization M_0 for the application of another RF pulse; the size of the longitudinal magnetization can be determined by substituting $t = TR$ in Equation 2.5. At this moment, the detected MR signal does not contain spatial information, as the signal's position within the sample cannot be deduced.

2.7 Position encoding

In general terms, a 3D image can be considered as a stack of 2D images. In the realm of MRI, the 2D images are termed *slices*. Each slice consists of a 2D *matrix* (the in-plane matrix). For convenience, the normal vector of the slices is set the direction of the z -axis here.

For position encoding, the dependency of the Larmor frequency on the strength of the magnetic field (see Equation 2.1) can be exploited: if a second, additional magnetic field is applied, which has a spatial variation in the form of a *gradient* in field strength, then a gradient in Larmor frequency is produced. Usually, a linear gradient is applied, i.e. the rate of change over space is a constant:

$$\frac{\partial B_z}{\partial z} = G_z \quad (2.7)$$

Hence, the net magnetic field now depends on the z -position:

$$B(z) = B_0 + z \cdot G_z \quad (2.8)$$

together with Equation 2.1 follows:

$$\omega(z) = \gamma(B_0 + z \cdot G_z) \quad (2.9)$$

In other words, the Larmor frequency is now dependent on the z -position. If an RF pulse centered around a certain Larmor frequency is applied in the presence of this gradient field, (which for convenience was set in z -direction), then only the spins around the *selected* Larmor frequency and thus a specific *location* will be excited.

The MR signal at this moment is a mix from all signals in the *slice*, since the region with constant net magnetic field is an orthogonal plane in the x - y -direction. For enabling in-plane position encoding, it is necessary to apply two further magnetic field gradients in x and y direction. The basic idea is to use two additional gradients to manipulate the MR signal in a way whereby it is possible to sample all *spatial frequencies* contained in the 2D slice.

Importantly, these two gradients are applied *after* the excitation pulse which initially perturbs the system. The 2D spectra containing the spatial frequencies are collectively referred to as *k-space*.

One of these gradients (set in y -direction here) introduces a phase difference of the spin precession. Crucially, this phase difference of the precession depends on the position along the y -axis, hence this gradient is known as *phase encoding* gradient. Depending on the y -position, the precession is either sped up or slowed down.

At the same time, on the x -axis, a frequency-encoding gradient is applied. This gradient changes the signal frequency depending on the location along the x -axis. Using both gradients, it is possible to sample the spatial frequencies contained in the 2D slice; in the simplest case one spatial frequency (and hence element of the k -space) at a time. After collection of all spatial frequencies, a 2D Fourier-Transformation is applied to the k -space to reconstruct the image[12, page 127].

2.8 T_1 and T_2 weighted images

Two main types of MRI images were acquired for my thesis: T_1 and T_2 weighted images. The former are used as anatomical reference (e.g. for spatial normalization of the brains into a common group space), the latter for functional images measuring brain dynamics (see [Section 3 on the next page](#)). In general, tissues with long T_1 time give a weak MR signal, while tissues with long T_2 time give a high MR signal[12, page 32].

It should be noted that both T_1 and T_2 weighted images profit from a higher external magnetic field strength B_0 , since the signal to noise ratio improves for higher field strengths and other factors (regarding the chemical environment of the spins) also contribute to an improvement at higher field strengths[12, page 168]. A qualitative comparison between the different field strengths for T_2 weighted images can be found in [Figure 3.1 on page 15](#).

Chapter 3

Functional MRI

3.1 Blood as contrast agent for MRI

The brain is traversed by a dense network of blood vessels, which provide it with important resources such as oxygen, glucose and various nutrients. At the same time, metabolic waste products have to be transported away (e.g. carbon dioxide, lactic acid etc.). The oxygen transport system has special properties, which can be exploited for functional MRI (fMRI), making it possible to visualize *in-vivo* changes of brain function. In the finest capillaries of the lungs, haemoglobin molecules residing in red blood cells bind oxygen. Hence the haemoglobin molecules become *oxygenated*. After releasing the oxygen at the target, the oxygenated blood can release the bound oxygen molecules and become *deoxygenated*. Importantly, the haemoglobin molecules change their magnetic properties depending on their *oxygenation state*. This effect was already discovered in the first half of the 20th century[13]. While oxygenated haemoglobin was demonstrated to be diamagnetic¹, deoxyhaemoglobin was shown to be paramagnetic². The different magnetic properties of the haemoglobin have an impact on measurements of magnetic resonance, because the paramagnetic deoxyhaemoglobin causes small local field inhomogeneities[7, page 90]. The magnetic inhomogeneities cause a decrease of the apparent transverse relaxation time T_2^* (see Section 2.5) [14]. In other words, the oxygenation level of the haemoglobin drives the MR signal as haemoglobin is a natural *contrast agent* for MR techniques. For this reason, the signal had been termed later on as blood oxygen-level dependent (BOLD) signal. Less than a decade after the discovery of the *in vitro* MR properties of haemoglobin, it was possible to perform a *in vivo*³ measurement of the BOLD signal [15].

¹diamagnetic materials generate a magnetic field in opposition to the external field and do not become magnetized

²paramagnetic materials enhance the external field and become magnetized as long as an external field is present

³the measurement took place in a rat's brain

3.2 Biophysics of the BOLD signal

To understand the origin of the BOLD signal in neural systems *in vivo*, the properties and mechanics of the cerebral blood circulation system have to be taken into account, in particular, the non-stationarity of blood flow: Blood vessels are able to increase their diameter, resulting in a local change (increase) of cerebral blood flow. As shown over a century ago[16], the dynamics of blood flow appear to be closely linked to the functional activity of the neural tissue. An increase in functional activity results in higher metabolic rates and thus a higher demand for blood supply, which is then met by enlarging the local vessels and thereby the blood flow. However, the underlying link between physiological mechanisms and the local neural activity are not yet fully understood [17]. In a general sense, it can be stated that neural activity is followed by a complex interplay between cerebral blood flow, cerebral blood volume and cerebral metabolic rate of oxygen consumption[18].

Interestingly, oxygen consumption increases in a disproportionally *smaller* rate than the cerebral blood flow [19]. In other words, the vascular response to increased local neural activity (often called the haemodynamic⁴ response) makes *more oxygenated* blood available, resulting in an *increase* of the BOLD signal.

The detailed mechanism of this *overcompensation* in oxygenated blood remains not fully understood to this date [20]. In contrast to the overcompensation of oxygenated blood, the glucose consumption levels are in line with the increase in supply [21]. Therefore, the surplus of oxygenated blood can be explained in two ways: it is possible that the demand is mainly driven by glucose metabolism. Since the ratio of oxygen and glucose is fixed and proportionally more glucose is extracted (as compared to the oxygen), it would be expected that the levels of oxygenated blood rise [20, 22]. Alternatively it is possible that the surplus arises from an inefficient delivery process of oxygen. In this scenario, the overcompensatory changes in blood flow are necessary to drive comparably smaller changes in the oxygen metabolic rate [23].

Furthermore, it should be mentioned that it remains elusive which *type* of neural activity drives the BOLD effect. It has been assumed that an increase of activity in local excitatory neuronal circuits causes an increase in BOLD signal. Recently, this assumption has been verified empirically using an optogenetic approach[24]. In brief, *in vivo* excitatory neurons were genetically modified to fire whenever they are illuminated with light of a certain frequency. Strikingly, the BOLD signal corresponded very closely to the optically induced firing patterns of the excitatory neurons. However other factors and their interplay may also play a role (given that any *sufficient* factor which drives the BOLD signal is not automatically *necessary* for it), for instance changes in excitation and inhibition balances[17].

⁴the ancient greek word for “haemo” translates to “pertaining to blood”

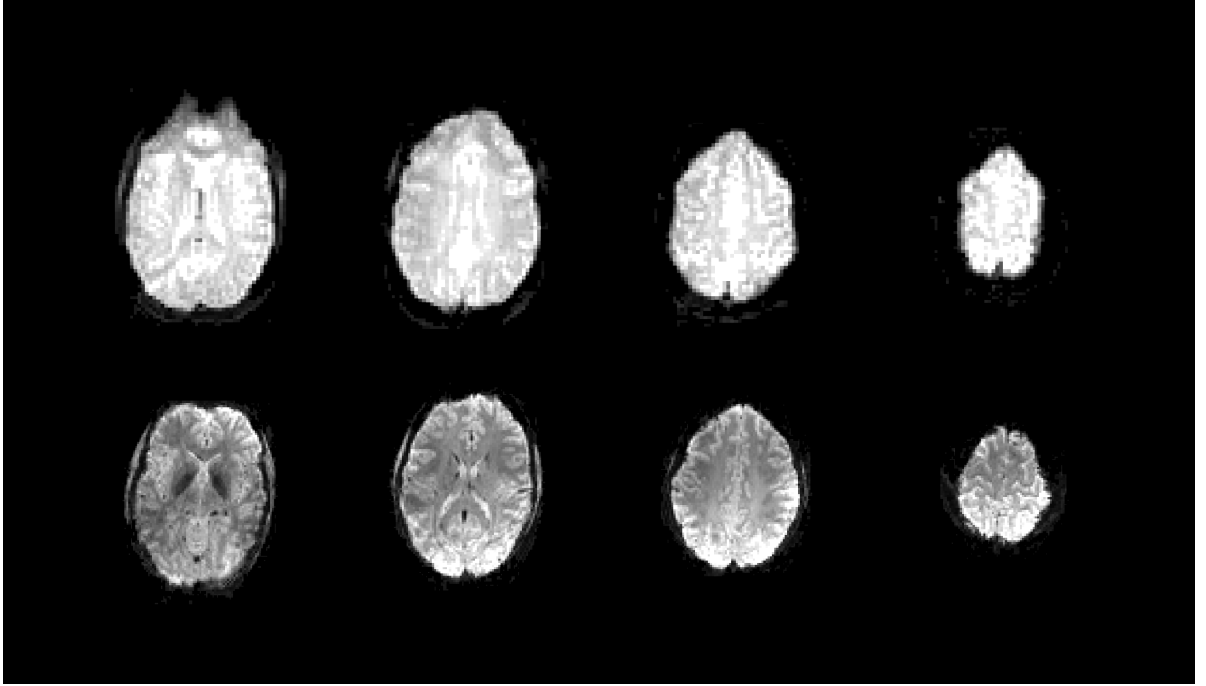


Figure 3.1: Comparison of a 3 Tesla BOLD fMRI scan (upper row) and a 7 Tesla BOLD fMRI scan (lower row) within the same subject.

3.3 Functional MRI recordings

In scanning practice, a BOLD image is acquired for every scanner repetition time TR. Typical values for the repetition times lie between two and three seconds, in this time a full 3D image of the BOLD signal of the participants head is acquired. The images consist of voxels, which have a resolution between 0.5mm and 4mm, depending on the MRI hardware and brain coverage. For a TR time of two seconds and whole brain coverage, typical resolutions are approximately 3mm isotropic for a 3 Tesla scanner and 2mm isotropic for a 7 Tesla scanner and about (see Figure 3.1). When only a Section of the brain is scanned and larger TRs ($> 3s$) are used, isotropic resolutions as high as 0.7mm may be achieved in a 7T environment. Recent developments, such as parallel imaging techniques[25] or compressed sensing[26] (where not the full k-space is sampled but only the parts of it) will likely make faster scanner repetition times or higher resolutions possible.

Over the timespan of an experiment, often a large number of 3D images are acquired, resulting in a 4D array $Y(x, y, z, t)$. The time series of the i -th voxel $Y(x_i, y_i, z_i, t)$ is displayed in Figure 3.2. The displayed voxel was located at the auditory cortex of a participant who tapped with his finger in synchronization to an auditory pacing sequence.

3.3.1 Correlational structure of the data

MRI recordings exhibit a correlational structure, both in spatial and in the temporal domain. Regarding the spatial domain, neighboring voxels are correlated to each other[27], which is likely due to the complex ways that the neuronal activity is measured by means of the underlying blood vasculature. Furthermore, specific preprocessing steps of the data (such as spatial

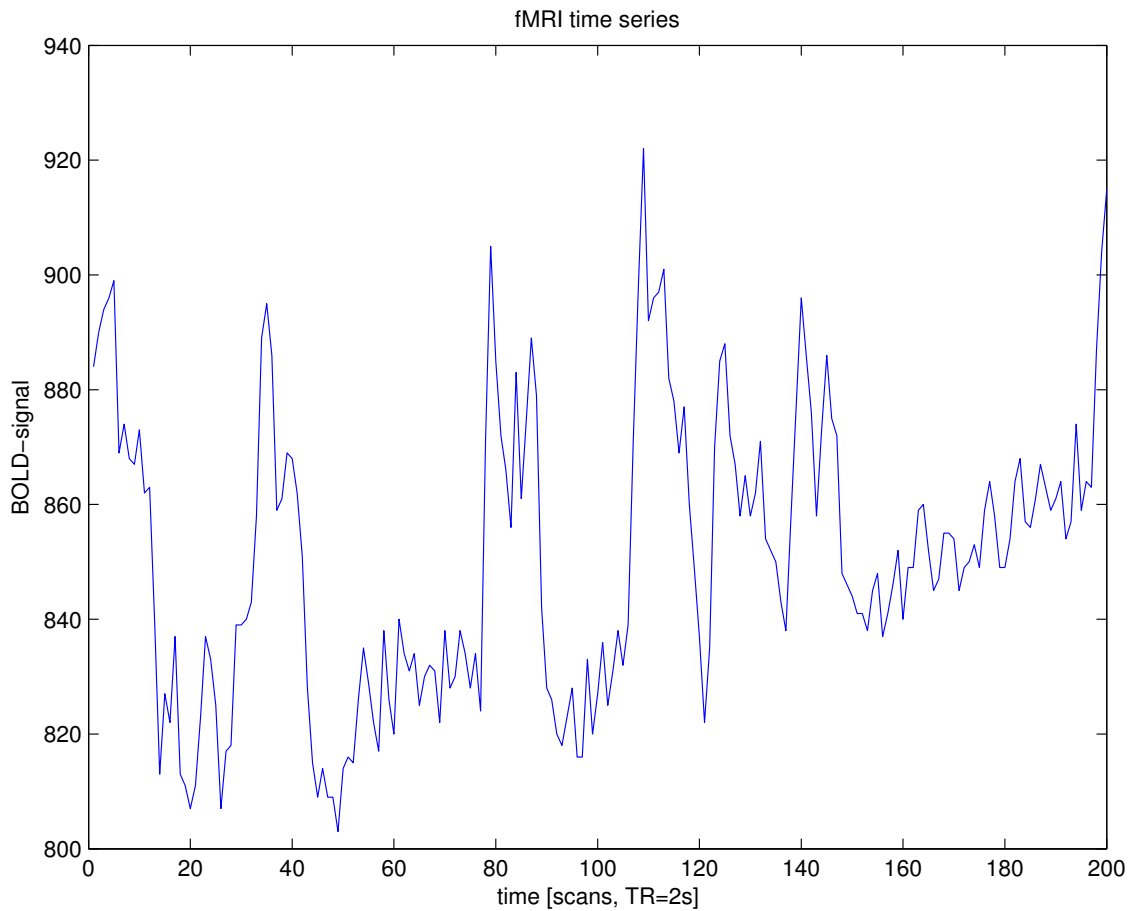


Figure 3.2: Exemplary BOLD time series of one single voxel

normalization to a common brain space or motion correction) may itself increase the correlation between neighboring voxels due to aliasing and interpolation effects.

FMRI data exhibits also a correlation in the temporal domain, subsequent scans show a considerable degree of correlation. The source of this temporal correlation lies in the underlying physiological fluctuations[28], which take place on much lower time scales than the temporal sampling rate.

Chapter 4

State of the art brain analysis

4.1 Human brain mapping

Over the last two decades, *brain-mapping* approaches have played a predominant role in human cognitive neuroscience. The principal idea behind brain mapping methods is to investigate *where* in the brain neuronal correlates of certain cognitive processes can be located¹.

The most straightforward and simple experimental design for brain mapping techniques utilizes a set of two experimental conditions, which are termed in the following A and B (see Figure 4.1). These conditions are presented repeatedly in a randomized fashion, and each presentation of such a stimulus is termed an *experimental trial*. In most cases, the experimental conditions are cued to the participants (often with a computer screen projection or headphones) and can involve task descriptions. For example, condition A could be a prompt to recall a memory from childhood and condition B be a prompt to recall an event from more recent time. Special emphasis is put on the design of the conditions so that that the *difference* between the conditions depicts the *desired* experimental factor, while other factors are minimized (so-called hidden variables). In the memory recall example above, the *type* of memory (short-term or long-term) could be regarded as the main experimental factor. However, this example may be prone to hidden variables such as the intensity of the memory recall or even the state of mind of the experimental subject, which covary with the condition in an unclear way.

Crucially, it is assumed that each experimental trial causes the subject to transition into one of the distinct measurable brain states α or β (see Figure 4.1), which are reflected in the recorded functional imaging data to a certain extent. Furthermore, it is generally assumed that the brain response of all experimental trials *within* a condition have highly comparable neuronal fingerprints, in other words that the brain responses of one condition are assumed to be similar to each other.

After data collection, the changes in the functional imaging data between the two brain states are computed. For this, a large variety of different statistical methods and analysis techniques

¹under a locationist's assumption, where brain function is no holistic but local

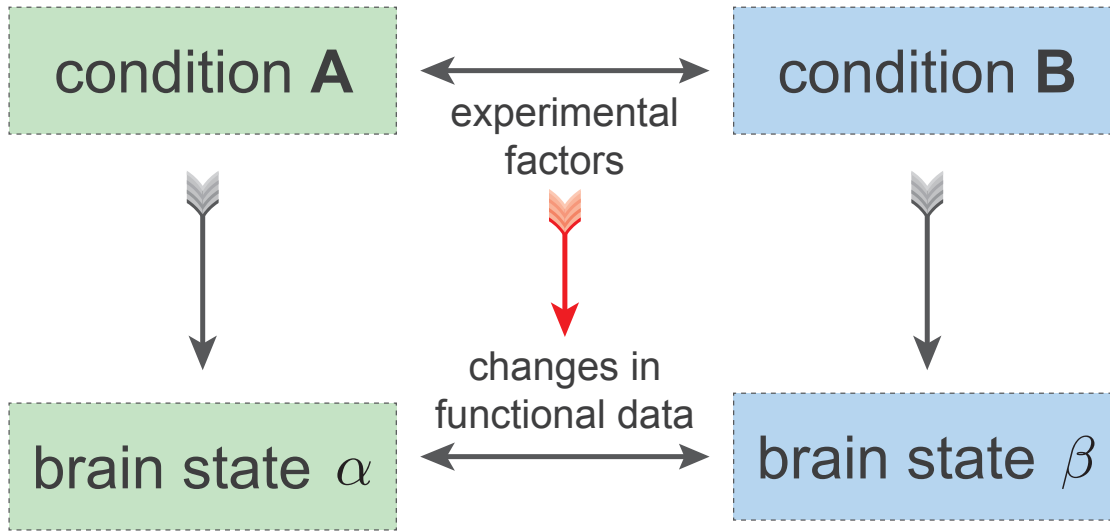


Figure 4.1: Experimental rationale for brain mapping studies

can be employed. Two kinds of such statistical methods form the centerpiece of this thesis; both are based on machine learning techniques and aim to visualize the locations containing information about task-related neuronal processing.

Ultimately, scientific inference is accomplished by causally linking the statistical changes of the functional brain image data to the variation of the experimental factors. To return to the example of a memory study described above, regions showing different responses to long-term versus short-term memories may be localized in the parietal cortex and the hippocampus of the test subjects, hence it may be concluded that these regions would be involved in encoding the *type* of memory.

4.2 Overview of fMRI analysis methods

In general terms, fMRI analysis methods can be divided into two classes. The first includes methods that analyze each voxel *independently* from each other, i.e. by treating each voxel as isolated and neglecting interactions between voxels. Hence, this class of analysis method is termed *univariate* analysis. Due to their conceptual simplicity and accessibility², univariate analysis methods have played a dominant role in the domain of imaging-based cognitive neuroscience. The second class of methods does regard interactions between voxels by analyzing *multiple* voxels *simultaneously*. Therefore, these methods are referred to as *multivariate* analysis techniques. The two analysis techniques used in this thesis (searchlight decoding and feature weight mapping) are both members of the second class, i.e. they are both multivariate. In the following, I will provide a brief overview of the most important univariate method, the *general linear model* (GLM), and will also introduce the most important and influential multivariate pattern analysis methods that have been used in fMRI research. It should be noted that to this day there exists a multitude of different analysis methods for fMRI data, and it is out of the scope of this thesis to give a complete review over all methods. A more general review

²implementations of the most common univariate analysis method, the general linear model, have been made available publicly in software packages such as SPM[29], FSL[30] and AFNI[31].

about current analysis algorithms can be found elsewhere [32], providing a larger frame for the methods discussed here.

4.3 The general linear model

The best-known member member of univariate analysis methods in imaging neuroscience is the general linear model (GLM). The earliest explicit usage of the GLM for imaging data dates back to 1995[33]. In a nutshell, the idea of the GLM approach is to find the best possible fit of an a-priori *generated* time course into the *measured* voxel-wise BOLD time course [32].

4.3.1 Mathematical formulation of the GLM

Assuming t time points, n voxels and m explanatory variables (experimental conditions and other variables of interest), the general linear model approach can be formulated as[34]

$$Y = X \cdot \beta + \epsilon \quad (4.1)$$

where Y is a $t \times n$ matrix containing the generated time-course data of all voxels and X is a $t \times m$ matrix containing the expected time courses of activity corresponding to each explanatory variable (which are generated by convolving the onset times with the hemodynamic mass response function, resulting in a simplistic time course as depicted in Figure 4.2). In other words, each column of X represents an explanatory variable. Most commonly, X is referred to as the *design matrix*. The matrix β of size $m \times n$ contains the unknown and to be estimated *weight* or *scaling* parameters, where each row in β corresponds to one explanatory variable. The error term (or model residual) ϵ of size $1 \times n$ contains everything that has not been captured by X and is assumed to be identically normally distributed.

The least squares *estimates* of β , commonly denoted as $\hat{\beta}$ is given by[35]

$$X^T X \cdot \hat{\beta} = X^T Y \quad (4.2)$$

Multiplication of both sides of Equation 4.2 with $(X^T X)^{-1}$ then solves for $\hat{\beta}$:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (4.3)$$

The absolute size of $\hat{\beta}$ usually is identified as the *level of activity*; large values for $\hat{\beta}$ reflect a high level of activity and low values for $\hat{\beta}$ a low level. The statistical analysis is carried out on the estimates $\hat{\beta}$, and in summary aim to detect how likely certain levels of activity arise by chance.

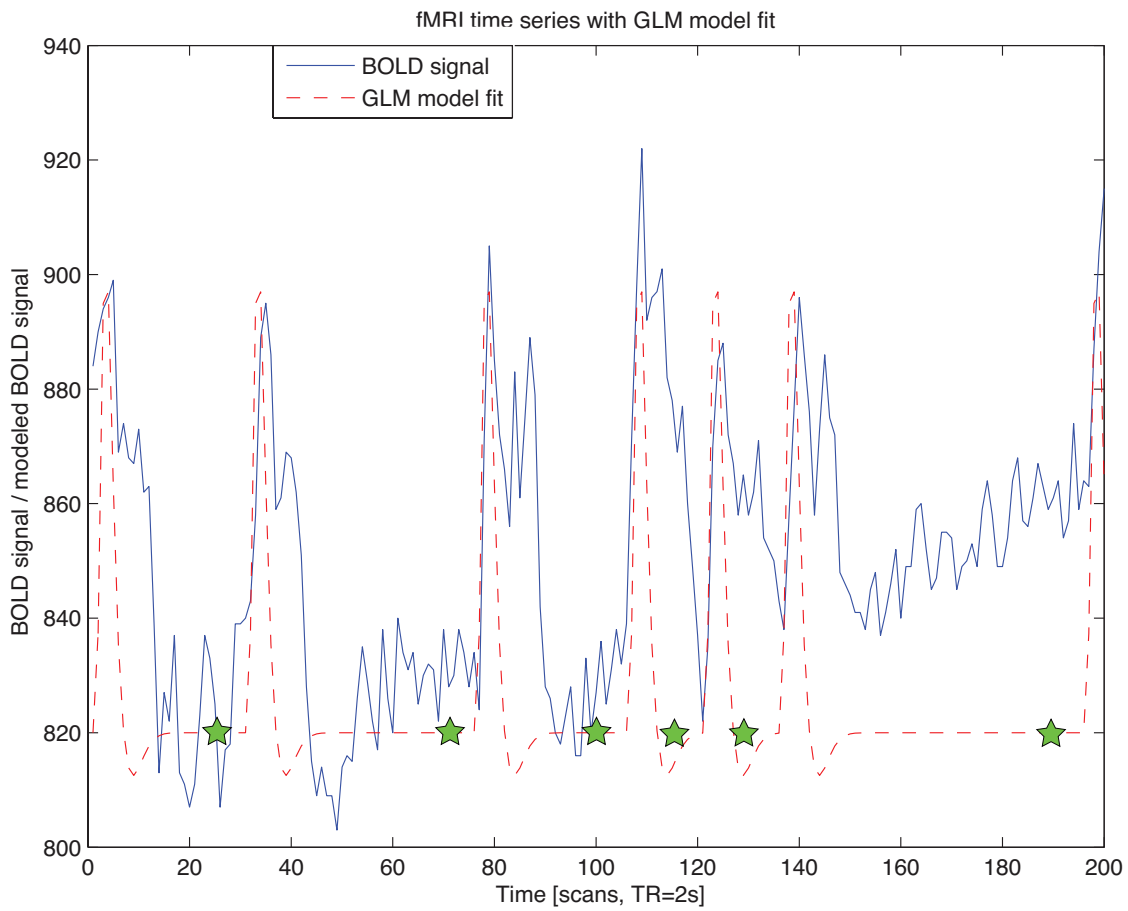


Figure 4.2: BOLD time series of one single voxel (blue) with a GLM data fit (red dashed). The data fit is derived by convolution of the onset times of the experimental condition (the onsets are depicted as green stars) with a generic haemodynamic response function

4.3.2 Criticism of GLM methods

Despite its simplicity, there are several points of criticism to the GLM approach. As the intention of this thesis is not to broadly exercise critique on univariate methods, I will limit myself to give a short summary of the three most important concerns. Firstly, the GLM approach is based on the idea of the logic of *cognitive subtraction*. According to this idea, cognitive tasks are *additive*, in other words a cognitive process is inset into another task or resting fluctuations without changing the latter (hence this assumption is termed “pure insertion”). To say the least, this assumption is extremely crude and fails at the *fundamentally nonlinear* nature by which most brain processes are governed[17]. Secondly, the generated data fits are not adequate. For instance, a large part of information is still found in the residuals ϵ of the GLM model, it has been shown that even after regressing out the experimental design, information about task-specific networks can be present[36]. Lastly, the human brain inherently processes information in local networks which interact together, often over large spatial scales. The importance of neuronal *communication* hence can hardly be understated. However, the GLM method does not explicitly consider neuronal communication processes. In light of this, it can be stated that the essentially multivariate nature of brain function is not taken into account by univariate models.

4.4 Multivariate pattern analysis

Multivariate pattern analysis (MVPA) methods adopt a different approach than the univariate method described above. The basic idea is to capture and analyze large-scale *patterns* of neuronal activity, which are spread over *multiple* voxels[32]. MVPA techniques have found their way into neuroimaging data analysis since the beginning of the 1990s when they were first applied to positron emission tomography (PET) data [37, 38]. In these earliest applications, the main focus of application was clinical, such as the diagnosis of brain diseases. Later on, when MVPA techniques were applied to task designs of cognitive neuroscience studies, they became increasingly widespread[39, 40, 41, 42]. Presumably, today's popularity of MVPA methods partially stems from the debate surrounding studies in the context of reading out unconscious thought [43, 32].

4.4.1 MVPA in neuroimaging

There are three main varieties of pattern-based methods for neuroimaging using fMRI: *decoding*, *regression* and *encoding*.

The most widely adopted approach is pattern classifier *decoding* [44]. Decoding techniques assume that different classes (categories) of experimental conditions or stimuli evoke different brain states (see Figure 4.1). Most crucially, it is assumed that diverging patterns of brain activation are produced by the different stimuli classes. This gives rise to the usage of pattern classifier algorithms, which make it possible to *categorize* the different brain response patterns and to *predict* their corresponding experimental category. In other words, pattern

classification methods use information from a pattern of brain activation to infer the task or state in which the brain is engaged [45][32]. At this point, it should be mentioned that both of the information mapping techniques utilized in my thesis are from this category of MVPA techniques (decoding-based) as they both implement classification.

The other two varieties of MVPA techniques, regression and encoding, are reviewed in more detail at the end of this Section (see 4.4.5 on page 27).

4.4.2 Classifiers

Decoding techniques rely on the principle of classification. The most simple classifier is a function that partitions a set of *data* into two classes [46]. The data space Y is of the dimension $t \times n$, where t is the number of examples $\vec{y} \in Y$ (also known as observations), each of which consists of n features $\vec{y} = [y_1, y_2, \dots, y_n]$.

As an illustrative example, let Y be the data space of weather recordings of some region. Y consists of the weather recordings of n weather stations, each of which recorded weather data at t time points. An example $\vec{y} \in Y$ hence corresponds to the weather measurements of all weather stations for a given time point t . The individual elements of \vec{y} (i.e. the features) are then the data value of a single weather station at the given time t .

Assuming the most simple case where only two *classes* of data points exist, the output space of the classifier are the labels $Z = \{-1, 1\}$. Crucially, there exists a distribution $D : Z \times Y$. Using this notation, we can define a classifier as function $f(\vec{y}) = z$, which maps an example $\vec{y} \in Y$ onto a label $z \in Z$. For this, the classifier first has to be trained on a subset of Y . The training (or design) of the classifier yields a *decision boundary*. Once this boundary has been derived, the classifier can be used to predict the labels of an unseen subset of Y in order to assess the *generalizability* of the learned relationship between features and labels. Classifiers differ in the way the decision boundary (often termed as hyperplane) is derived[32].

Linear classifiers achieve the classification decision by computing the dot product between the example \vec{y} with a weight vector \vec{w} , the latter of which is derived by the training. If the product $\vec{w} \cdot \vec{y} > c$, the classifier decides for the first class, while the classifier determines the second class for $\vec{w} \cdot \vec{y} < c$, where c is a constant (all under the assumption of a two-class design).

By usage of the dot product, each feature influences the decision only by its weight (interactions across features have no direct influence) [47]. The following linear classifiers have been used extensively in the context of fMRI decoding:

- Fisher's linear discriminant analysis
- Linear support vector machines
- Pattern correlation classifiers

Non-linear classifiers, on the other hand, allow interaction between the features. The most widely used non-linear classifiers in the context of fMRI are

- Gaussian naive Bayes
- non-linear support vector machines (Gaussian radial basis or polynomial kernels)
- Artificial neural networks

This group of classifiers provides a more flexible decision boundary, however they are also more prone to overfit the training data, i.e. adjusting the decision boundary to aspects of the data that are actual noise[32](see Figure 4.3).

The different types of classifiers adopted for fMRI data have been compared quantitatively in literature[47, 48]. It was found that linear classifiers appear to perform comparably to non-linear classifiers. Linear classifiers also exhibit a better stability and interpretability of the results[44], while lowering the risk of overfitting (see Figure 4.3). Moreover it should be noted that linear classifiers often have a smaller computational cost.

4.4.3 Feature Selection

The data matrix Y in fMRI data is of the size $t \times n$, where n is the number of voxels and t the number of examples (e.g. experimental trials); typically n is very large (about 50000 for 3 Tesla data and 500000 or more for 7 Tesla data, see Figure 3.1), while t is small (between 10 to 100). As t is comparably very low, there exists a high chance that the classifier overfits the data[49]. Therefore, the *dimensionality* of the feature space n most commonly is *reduced*. The reduction of the feature space in the case of fMRI usually increases performance, for instance it appears straightforward that it may well be beneficial to exclude a voxel containing both negligible levels of information and high levels of noise. There exist a variety of different strategies for the selection of features. In the following I will briefly describe the most common approaches used in the context of fMRI. Two of these approaches are especially important for this thesis: the searchlight approach and the dimensionality reduction approach using matrix decompositions.

4.4.3.1 Wrapper methods

For distinguishing important from less informative features (voxels), so-called wrapper methods can be used [50]. In a nutshell, wrapper methods treat the classifier as a black box and score subsets $Y' \subset Y$ of the data space by comparing the predictive power of each subset. Ultimately, the subset with maximum predictive power is attempted to be found.

There exist two types of wrapper methods; *backward elimination* and *forward selection* [51]. Backward elimination techniques such as the *recursive feature elimination* (RFE) approach start with a large set of features (for instance, the entire space Y) and compute a metric for the

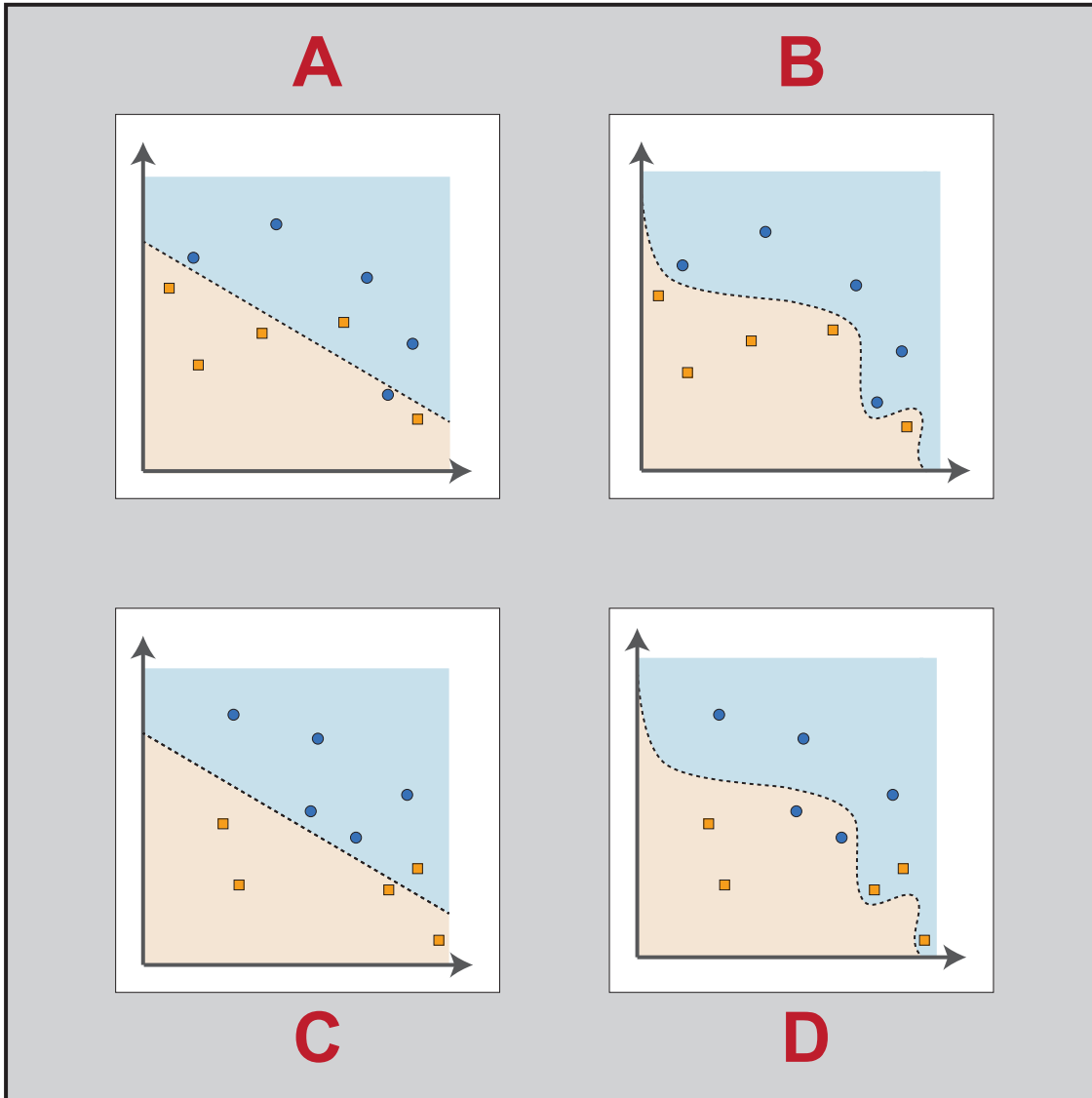


Figure 4.3: Comparison between linear and non-linear classifiers. The classifier labels data points within the orange area as belonging to class 1, while data points within the light blue areas are classified as class 2. The actual class (ground truth) of each data point is represented by the filling and shape of the data points; orange squares are data points of class 1 and blue circles are data points of class 2. **(A)** A linear classifier is trained on a data set. The resulting decision plane does not separate the classes perfectly. **(B)** In contrast, a non-linear classifier (in this case a higher order polynomial) may find a decision boundary with an error rate of zero. However, an overfitting problem could arise from this decision boundary, as it reflects potential noise in the data. **(C)** The decision plane derived in A is now used to classify a test set of new, unseen data. Since the error rate in the classification of the test set is low, the generalizability of the classifier is regarded as high. **(D)** When the decision boundary derived in B is used for new, unseen data, it is possible that the classification result is prone to error. This is due to the overfitting that took place in the classifier's training (in B).

features. This allows a *ranking* of the features. In an iterative way, the most irrelevant features (the ones with lowest ranks) are discarded until a set of the most discriminative features is distilled, which yields a discriminating map of voxels[52].

Examples for *forward selection* techniques include genetic algorithms [53]. In an similar fashion to evolutionary processes, features are exchanged (*mutation*) and the discriminative power of the new set is evaluated (*selection*). Over many iterations (*generations*), a maximally discriminative set of voxels can be worked out [54]. The crucial question when using wrapper methods is the definition of the stopping criterion. Generally, either a criterion based on the change rate of performance between subsequent iterations is used, or the best feature set is selected post-hoc as the one with the highest performance [52, 32].

4.4.3.2 Region of interest analysis

A simplistic solution for narrowing down the number of features is to perform a *spatial selection*, i.e. to select a region of interest (ROI)[32]. Hence, the reduced data space is a subset of the original data space $Y' \subset Y$. In general, ROIs can be defined by two different approaches; either by an inclusion criterion based on *functional* MRI data, or alternatively by *anatomical* (or structural) means [55]. A very common approach for determining functional ROIs are so called functional localizer scans. The localizer scans are run before or after the actual experiment and consist of a paradigm (the localizer task), that serves to *functionally map out* an area of interest. Usually, an univariate inclusion criterion in the form of a GLM is used. In other words, voxels showing large levels of activity (large β -estimates, see Section 4.3) in the localizer task are selected to be part of the ROI [40, 41, 56, 57]. It should be noted that univariate methods for feature selection may very well discard voxels that actually do encode information in regards to the localizer task, but the information is not reflected in the sole magnitude of the univariate β -estimates [58]. It is possible that, this issue is mitigated if multivariate approaches are used for the feature selection itself, such as the searchlight approach or wrapper methods (see Section 4.4.3.4 and Section 4.4.3.1).

On the other side, anatomically defined ROIs also can provide interesting insights, especially if connections between structure and function are of interest. There are many routes to the derivation of anatomical regions of interest. One possibility for defining structural ROIs is the usage of probabilistic atlases, which indicate the *probability* of membership of a given voxel coordinate to distinct brain areas[59, 60]. The golden road to a perfect anatomical parcellation into regions of interest, however, would be a *microstructural* parcellation based on the cyto- and myeloarchitecture of the cortex[61]. This would allow the tailoring of anatomical ROIs for each individual subject. Most importantly, this approach would overcome the limitations of spatial normalization into a common inter-subject brain space (see Section 8.3 on page 63). At present, however, there is no implementation of “in vivo Brodmann mapping” available that automatically parcellates the cortex into ROIs based on anatomical fine structure measured by MRI[32].

4.4.3.3 Dimensionality reduction of the feature space

As their name already suggests, *dimensionality reduction* approaches transform the data space Y to a space Y^* with fewer dimensions. Most commonly in the context of fMRI, the number of examples is kept unchanged while the number of features is reduced. There exist numerous approaches for doing so, for instance clustering, basic linear (and other) transformations or convolutions[51, 32]. In this thesis, I will focus on one class of linear transformations, namely the principal component analysis (PCA). The PCA approach is one of the most commonly practiced dimensionality reduction methods used for functional MRI data [37, 62, 42], particularly in regards to a subsequent application of pattern classification methods within the dimensionality reduced space Y^* . Pattern classification results can then be projected back from the PCA space Y^* into the voxel space Y for visualization and voxel-wise statistics[4]. One of the information mapping techniques (the *feature weight mapping* method) described in my thesis builds on this dimensionality reduction using PCA and also involves a back-projection into the voxel space.

4.4.3.4 Searchlight approach

Searchlight methods traverse the three spatial dimensions of functional MRI data in a 'scanning' fashion [63] and select a (commonly spherical) neighborhood of voxels as features for classification. Typical spherical diameters are in the range of about one centimeter. In other words, as in the ROI approach, the reduced data space is a subset of the original data space. However, the searchlight procedure is repeated for every location. It is then common to map the classification result back into each investigated location, which yields a map of classification results if repeated over all locations.

Most commonly, the searchlight is *volumetric*, however recently, this approach has been adopted for cortical *surfaces* [1]. The searchlight method yields a high performance in classification and gives spatially unbiased estimates (as there is no spatial hypothesis) of *local* information content. As only local environments are taken into account, the searchlight approach does not capture large-scale interactions and jointly encoded information of distinct brain areas[32].

4.4.4 Cross-validation

In order to estimate the generalization error of the classification, a cross-validation often is applied[64, page 483]. For this, the data set Y is split into a training and a test set. Most commonly in the context of fMRI classification and a two-class design, a *leave-one-out* cross-validation scheme is applied. Given a data matrix Y of the size $t \times n$, a total number of $t/2$ cross-validations is used. Each time, the classifier is trained on a subset of size $t_{tr} \times n$, where $t_{tr} = t - 2$, in other words one single training point from each class is excluded for the training. The remaining two examples (one of each class) are used as test sets. In the case of fMRI, the examples are not chosen at random but rather in a subsequent fashion.

The resulting estimate of accuracy is the number of correctly predicted labels from all cross-validation folds divided by the total number of predicted labels.

4.4.5 Pattern analysis methods beyond classification

In the following, I will briefly discuss the two remaining main categories of pattern classification approaches; regression and encoding.

4.4.5.1 Regression

Instead of *classifying* the fMRI data into different classes using a discrete output label space (in a two-class design $Z = \{-1, 1\}$), it is also possible to *regress* the data to a *continuous* response variable of a given interval, e.g. $Z = [0, 1]$. The response variable Z usually is not an fMRI measurement, but likewise is a response variable provided by the subject directly (for instance a continuous task response such as reporting how positive the current thoughts are). Alternatively, Z may be a measurement from another modality, such as heartbeat, respiratory or electrophysiological signals.

Linear regression allows the determination of a set of voxels, which best describes the response variable Z . The best estimate of the response variable Z then is given by

$$\hat{Z} = Y\beta \tag{4.4}$$

where Y was the $t \times n$ matrix containing the time-course data of all voxels and a total of n β -coefficients are estimated. The approach looks similar to the GLM approach (see Section 4.3.1 on page 19), however in this case, *all voxels* simultaneously regress a response variable, whereas in the GLM approach a linear combination of a generated BOLD signal (the design matrix) regresses the actual BOLD signal of a single voxel.

Similarly to GLM, the standard method for coefficient estimation is an ordinary least squares approach. Alternatively, a *regularization* can be introduced[65], by applying a constraint on the β -coefficients of the regression model. For instance, L_1 norm regularizations can be applied which introduce a constraint on the sum of the absolute values of the coefficients[32]. The LASSO method [66] uses such a constraint and was shown to result in sparse solutions. Further regularization constraints use L_2 norm regularizations[67] (introducing a constraint on the sum of the square values of the coefficients) or combination of both [68, 65, 69]. Alternatively, multivariate regression can also be performed using support vector regression[70].

Note that multivariate regression is strongly underdetermined, as in general there are vastly more voxels than time points in fMRI data. Therefore, the results of this procedure depend heavily on the choice of constraints. For instance, sparsity constraints may be problematic as it is not clear to what extent brain function really is sparse[32].

4.4.5.2 Encoding

Arguably, one of the weak points of classification or regression methods is the limitation regarding the neurophysiological *interpretability* of the data. While these methods allow one to claim that there *exists* a difference in the evoked response patterns of two conditions and possibly also, *where* this difference is located within the *feature space*, it is factually impossible to gain insight into the underlying mechanisms driving the formation of these patterns. In other words, classification and regression methods treat the underlying brain function as a “black box”. However, the understanding of such neuronal mechanisms is arguably a centerpiece of any meaningful theory of brain function.

Encoding approaches are complementary and attempt to fill the gap described above. In contrast to decoding approaches, encoding methods *generate* the BOLD response given an experimental stimulus. Unlike the GLM approach (see Section 4.3), which commonly only takes the *onset* times and a generic hemodynamic response function to generate a BOLD time course, encoding approaches use the *entire* experimental stimulus. Commonly, a number of features are extracted from the experimental stimuli (as the dimensionality of the experimental stimulus often is very high). For instance, given movies as experimental stimulus[71], information regarding the edges and corners could be extracted from the video directly and used as features. In a next step, a *relationship* between these stimulus features and the recorded BOLD signal is estimated (e.g. by using regression techniques as described in Section 4.4.5.1 on the preceding page). Given this model of brain response to specific experimental input, it is possible to predict the brain response for previously unseen experimental stimuli. Crucially, the quality of the model can be evaluated by a simple comparison, for instance by assessing how much of the variance in the data can be described by the model. Encoding approaches mostly find application in prediction of early sensory (mostly visual) areas [71, 72].

4.4.6 Criticism of MVPA methods

4.4.6.1 Decoding and regression

The main critique on pattern based techniques for fMRI analysis is commonly that it is only possible to determine *where* informative features for classification are located (e.g. by the usage of feature selection methods). From a neuroscientific point of view, the more interesting question would be *how* information and information processing are represented and carried out in the living human brain. However, classification- and regression-based methods can rather be described as “black box” models, and do not offer insight into the characteristics of neuronal representations.

4.4.6.2 Encoding

Encoding models do not have the above limitation of *interpretability*. On the other hand, encoding models have a conceptual weakness in the *construction* of the model, where features

are computed from the experimental stimuli. Presently, at best, these features are extracted from the experimental stimuli by “neurally inspired” mechanisms [72]. It remains unclear to what extent these mechanisms reflect the actual underlying neuronal computations - a problem shared by almost all models proposed in cognitive neuropsychology[32].

4.5 Activation vs. Information mapping

Over the last two decades, so-called *activation*-based methods have been predominant in imaging-based cognitive neuroscience. The idea behind activation based paradigms is to estimate the magnitude of a large-scale neuronal response to an experimental stimulus, and to compare these response magnitudes across experimental conditions. For this, most commonly, the general linear model (see Section 4.3) is used.

Information based approaches, on the other hand, attempt to quantify the mutual information between the experimental stimulus/condition and the measured brain signal[45]. For assessing the mutual information, multivariate methods (see Section 4.4) such as decoding methods are often employed.

Both approaches differ substantially in the underlying assumptions about the neuronal representations elicited by experimental stimulation. While in activation-based methods, the size (and variance) of the responses determines the statistical difference, the absolute response size does not necessarily have to be dissimilar to information based methods, as information may be present in the form of a relationship amongst voxels or pattern of voxels.

Commonly the overall signal is assumed to consist of two components. Firstly, a *smooth* component (which is spread out over a neighborhood of voxels) and secondly, a *fine-grained* component[45]. In activation based approaches, the *fine-grained* component is considered to be noise. Therefore, spatial filtering in the form of a Gaussian filter kernel is applied prior to the data analysis, which effectively removes the fine grained component. Usually, this smoothing drastically improves the GLM fit.

Information-based approaches, however, do not require spatial smoothing, as also the fine-grained component is suggested to contribute to the discriminability of the patterns. At higher resolutions or smaller voxel sizes, the smooth component typically is smaller, while the fine grained component is more pronounced[45]. Therefore, information-mapping techniques are often the preferred methods for analyzing high-resolution data.

In the following, I will present two of the most relevant information mapping techniques for the present thesis, namely searchlight decoding and information mapping using feature weights.

4.5.1 Searchlight decoding

As the searchlight approach was already introduced in Section 4.4.3.4 on page 26, I will only briefly summarize the logic behind information mapping using searchlight decoding. In essence, task-related neural information at every location in the brain is assessed by analyzing the signal patterns extracted from a spatial neighborhood (the searchlight) centered at each location of the brain[5]. Searchlight *decoding* can be regarded as a type of analysis strategy using the searchlight approach, which uses classification-based MVPA. Decoding involves training a classifier on a subset of the data and predicting the class labels of another, yet unseen subset of the data. Thereby, the generalizability of the classifier is assessed[44]. Usually, this procedure is repeated for different subsets of data, in other words a cross-validation procedure is applied (see Section 4.4.4 on page 26). The average percentage of correctly predicted labels, known as the *decoding accuracy*, is taken as an indicator of the information content of the searchlight volume. Customarily, the accuracy is mapped to the central voxel of the searchlight. The repetition of this procedure for all searchlight locations in the brain mask hence results in a three-dimensional accuracy map, which reflects the spatial distribution of information decodable from the functional brain images. It should be noted that in the context of neuroscientific studies, the decoding accuracies themselves are usually not of primary interest. Instead, the statistical significance of the decoding accuracy is of highest relevance[5].

4.5.2 Feature weight mapping

Alternatively, information maps can be computed without the spatial preselection of voxels applied in the searchlight approach, by training a (linear) classifier directly on whole brain fMRI data. This yields the *weight* vector \vec{w} (see Section 4.4.2 on page 22). Crucially, each component of this weight vector belongs to a feature dimension and indicates what *influence* this feature exerts in regard to the classification decision[?]. Since the feature space (i.e. the number of voxels) in fMRI data typically is very large, a dimensionality-reduction procedure is commonly applied before the classifier is trained (see Section 4.4.3.3 on page 26).

In analogy of the decoding accuracies for the searchlight decoding approach, the size of the feature weights are not of the primary interest, but rather their statistical significance.

Chapter 5

Statistical inference

5.1 Hypotheses testing

The term "statistical inference" refers to those methods and procedures, which allow one to draw *reliable conclusions* on the basis of experimental data. More precisely, statistical inference methods allow a selection of an *optimal explanation* from a set of possible explanations, provided experimental data. The explanations are commonly referred as *hypotheses*.

As illustrative example a data set can be considered which includes treatment reports of two groups of clinical patients, who all are suffering from the same disease. One group is given a new drug that is anticipated to treat the illness, while the other group is given a placebo medication. Statistical inference methods allow one to gain quantitative insights in data, effectively making it possible to draw conclusions about the curative effects of the new drug. In this example, the following sets of hypotheses appear plausible:

H_0 : The new drug *does not* have any curative effect in regards to the illness other than the placebo-effect¹

H_1 : The new drug does have healing effects and cures the illness

The hypothesis H_0 is known as the *null hypothesis*. The hypothesis H_1 is usually referred to as *alternative hypothesis*. The central idea behind scientific reasoning using frequentist[73] statistical inference is to estimate how likely is it to obtain the measures *given chance*, i.e. given that the null hypothesis H_0 actually is true. Most importantly, it is desired to avoid the scenario where the null hypothesis indeed was true, but was rejected on the false premises. The incorrect rejection of a true null hypothesis is known as *type I error* or *false positivity*. As the prerequisite of frequentist statistical inference is to minimize the occurrence of type I errors, a statistical threshold α is defined, indicating the probability for committing a type I error. Only if the probability for the null hypothesis H_0 being true is determined to be smaller than α , the null hypothesis H_0 is rejected. The threshold α is commonly denoted as *significance level*. Typical values for α are 0.05 or 0.01, i.e. the null hypothesis H_0 is rejected

¹in other words, H_0 can be formulated as: the new drug has the same curative effects as the placebo medication

	<i>In reality H_0 is true</i>	<i>In reality H_0 is false</i>
test <i>rejects</i> H_0	false positive type I error	true positive
test <i>doesn't reject</i> H_0	true negative	false negative type II error

Table 5.1: Relations between the truth or falseness of the null hypothesis H_0 and the outcomes of a test

if the probability of H_0 being true is smaller than 5% or 1%. The probability of H_0 being true is commonly denoted as p -value.

In order to decide what probability the null hypothesis H_0 has given data, the *null distribution* of the test statistic has to be assumed or alternatively computed empirically. This then allows to a determination of the probability of H_0 being true (i.e. the p -value) given the data and a decision of whether H_0 is rejected or not. If the form of the underlying null distribution is known, the problem is *parametric* and can be solved using parametric statistical inference. On the other hand, if there is no knowledge about the distribution, the problem requires nonparametric statistical inference methods[74, page 34]. In the following Sections, I will discuss the most important parametrical and non-parametrical statistical tests.

The reverse scenario of a type I error is also possible, where the alternative hypothesis H_1 was factually true but the null hypothesis H_0 is selected (i.e. the null hypothesis H_0 was *not* rejected). This decision is known as a *type II error* and commonly also denoted as a *false negative*. An overview over false positivity and negativity is provided in Table 5.1. The probability of a type II error is complementary to the *power* of the statistical test, which is defined as the probability of deciding on the alternative hypothesis H_1 when H_1 is in fact true[74, page 17]. Since in the normal experimental setting, the ground truth typically is inaccessible, it is not possible to conclusively compute the number of type II errors. In simulations where the ground truth is known, however, the computation of type II errors or power is possible².

Another measure for the quality of a statistical test is the precision (assuming the ground truth is known). The precision is defined as the ratio between *true* positives and *all* positives (i.e. true positives divided by the sum of true and false positives, see Table 5.1). Hence, the precision takes the value of the interval $[0, 1]$, indicating the fraction of positive tests that had indeed been true positives.

Furthermore, the sensitivity of a test can be computed, which is the number of true positives divided by the sum of true positives and false negatives (see Table 5.1). Hence the sensitivity also takes values of the interval $[0, 1]$. The sensitivity indicates what fraction of ground truth positives had been correctly identified as positives.

In general terms, frequentist statistical testing procedures always effectively incorporate a trade-off between the precision and the sensitivity: The higher the sensitivity, the lower the precision; the higher the precision of a test, the lower it's sensitivity. The issue is illustrated schematically in Figure 5.1.

²For this reason, the proposed statistical frameworks of my work are applied to a variety of artificial data in the form of simulations of brain activity data.

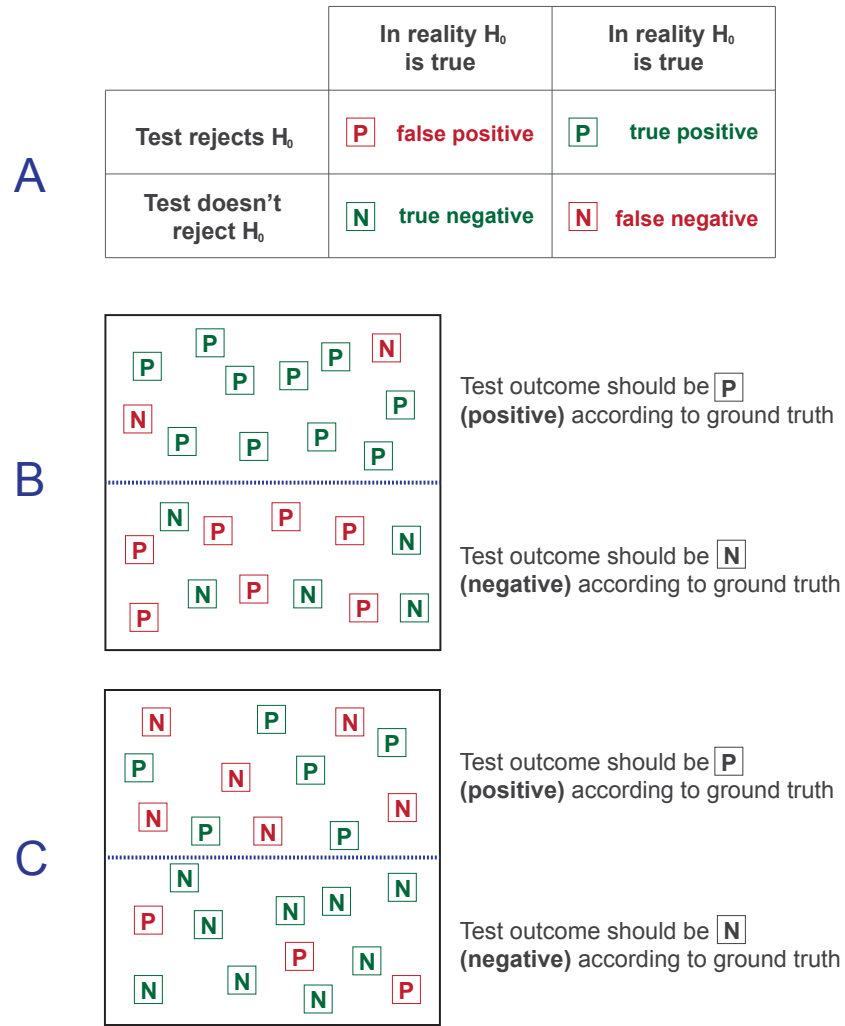


Figure 5.1: Trade-off between sensitivity and precision of statistical testing procedures. **(A)** Summary of the relations between the truth or falseness of the null hypothesis H_0 and the outcomes of a test, as in Table 5.1. However in here, four symbols are introduced in the form of the two letters P and N and two colors red and green. The P's indicate that the test result was *positive* (H_0 was rejected), the N's indicate *negative* test results (H_0 was not rejected). The colors indicate whether the outcome of the test was in accordance with the ground truth; green indicates that this was the case, red indicates a mismatch between the ground truth and the test result. **(B)** Outcome of a statistical test which has a high sensitivity but therefore a low precision. According to the ground truth, all tests in the upper half of the cube (separated by the blue dashed line) should have been labeled positive, while all tests in the lower half of the cube should have been labeled negative. As the sensitivity of the test is high, most tests in the upper cube are labeled correctly (i.e. as positive). On the other hand, the price for the high sensitivity is that the precision of the test is low, a large fraction of the tests in the lower half had been falsely labeled positive (instead of negative). **(C)** Outcome of a test featuring a lower sensitivity but therefore a higher precision. As before, the upper half of the cubes should have been labeled positive and the lower half negative, in accordance to the ground truth. Since the sensitivity of the test is low, many positives in the upper half remain undetected (i.e. labeled negative). In the lower half, however, most tests are labeled correctly as negatives, as the precision of the test is high.

5.2 Parametrical statistical inference

Parametrical test procedures imply certain assumptions on the *distribution* of the experimental data. For instance, parametric statistics impose assumptions on the *form* of the distribution from which the data points are drawn, for instance an underlying normal distribution. Furthermore, the parameters of this distribution are also assumed (or approximated). It can be shown that parametric tests are the most *powerful* statistical tests (i.e. they have the lowest probability for type II errors)[74, page 37], provided the above assumptions are true.

5.2.1 Z-test

The Z-test (or Gauss-test) is one of the simplest parametric tests and assumes that the observations x_1, \dots, x_N are drawn independently from a normal distribution. Furthermore, the standard deviation σ and mean μ_0 of the underlying normal distribution are assumed to be known.

The mean μ of the observation is given by:

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

The null hypothesis for Z-tests states that there is no effect, in other words the population mean μ_0 given *no effect* and the measured sample mean μ are identical:

$$H_0 : \mu = \mu_0$$

The Z-test statistic then can be formulated as[74, page 37]

$$Z = \frac{\sqrt{N}(\mu - \mu_0)}{\sigma} \quad (5.1)$$

The probability density of Z is:

$$f(Z) = \frac{e^{-\frac{(Z-\mu_0)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \quad (5.2)$$

The probability density $f(Z)$ is normalized:

$$\int_{-\infty}^{\infty} f(Z) dZ = 1 \quad (5.3)$$

The probability of an observation with the value $Z \geq Z_0$ is then given by the integral over the probability density $f(Z)$:

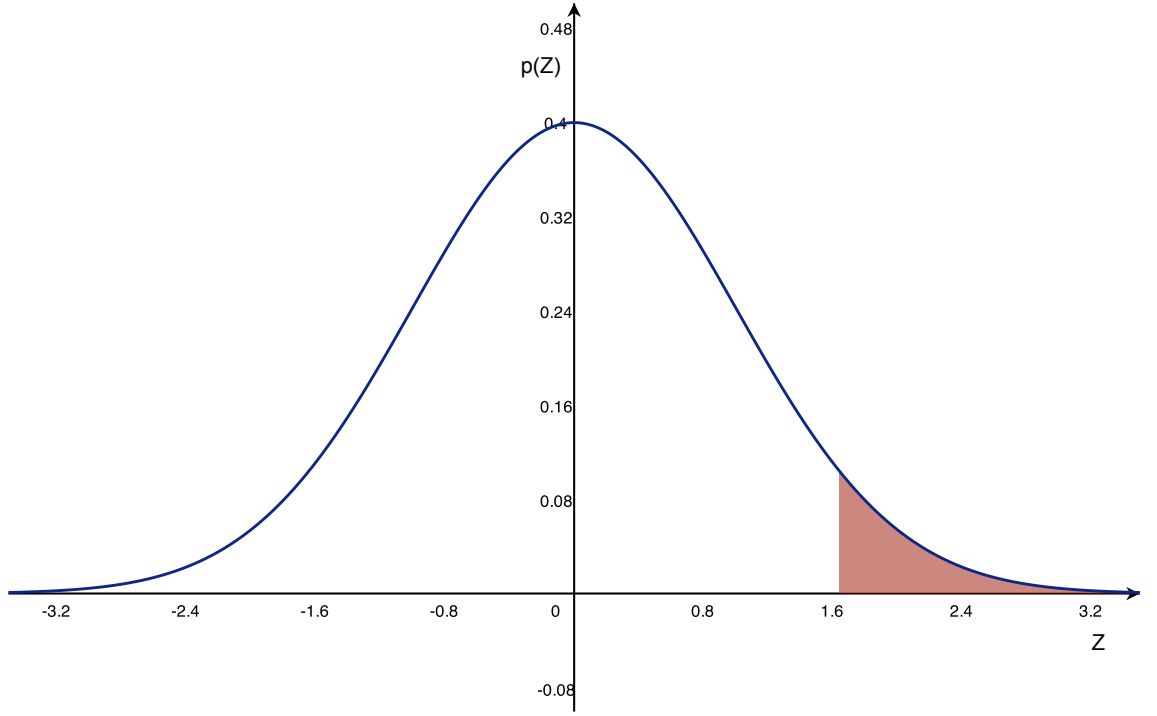


Figure 5.2: Z-distribution with the parameters $\mu_0 = 0$ and $\sigma = 1$. The probability for finding a Z-value of Z_0 or larger is given by the area under the probability density of Z . In this illustration this value is determined as $Z_0 = 1.645$. The probability of finding a Z value bigger than 1.654 then corresponds to the relationship between the red shaded area to the total area under the curve. In here, the red shaded area covers 5% of the total area, corresponding in a p -value of 0.05

$$Pr\{Z \geq Z_0\} = \int_{Z_0}^{\infty} f(Z) dZ$$

A visualization and geometrical interpretation of this derivation of probability given the null distribution are displayed in Figure 5.2.

5.2.2 T-test

Usually, the standard deviation σ of the underlying population is unknown. The student's T-test (or short T-test) resolves this lack of knowledge by substituting the standard deviation of the population σ in Equation 5.1 with an estimation of the sample variance $\hat{\sigma}$

$$\hat{\sigma} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2} \quad (5.4)$$

The test statistic then becomes[74, page 38]:

$$T = \sqrt{N} \frac{\mu - \mu_0}{\hat{\sigma}} = \frac{\sqrt{N}(\mu - \mu_0)}{\sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}} \quad (5.5)$$

In analogy to the Z-test, the probability of finding a value $T \geq T_0$ is derived by integration

over the probability density over T . The probability density of T can be derived[75] as:

$$f(T, \nu) = \sqrt{\frac{\nu}{2\pi\mu}} e^{\frac{-T^2\nu}{\sigma}} \quad (5.6)$$

where $\nu = N - 1$ is the degree of freedoms of the test.

The T-test is one of the most widely used statistical tests and assumes the data to be normally distributed and sampled from a continuous distribution. T-tests have also been employed for classification-based decoding methods to determine the *statistical significance* of decoding accuracies (which was defined as the percentage of correctly estimated labels). For this, the following null hypothesis is formulated:

H_0 : The mean decoding accuracy μ is identical to the chance level mean decoding accuracy $\mu_0 = 0.5$

Most commonly in the context of classification-based fMRI, the T-test is carried out on the group level, using the mean decoding accuracies from each individual subject as data points. In other words, given N_{sub} subjects, the mean decoding accuracies $x_1, \dots, x_{N_{sub}}$ are used for computing $\hat{\sigma}$ and μ_0 .

5.2.3 Binomial models

In the context of classification, there exists another way of parametrically deriving the null distribution of accuracies: the classifier can be modeled as a Bernoulli trial [47]. Here, the null hypothesis can be broadly defined as:

H_0 : There is no class information present in the data

This has as a consequence the need for the classifier in practice to *guess* the labels of the test set[5]. Let us assume that an already trained classifier is used to estimate N class labels of an unseen test set, containing N data points. Since the N data points are independent from each other, it is possible to compute a theoretical null distribution by assuming N independent Bernoulli trials. The number of correctly estimated labels of the N trials thus can be represented by the binomial random variable X , which depends on the parameters (N, p) . The probability mass function³ of X then is:

$$f_X(c) = \binom{N}{c} p^c (1-p)^{N-c} \quad (5.7)$$

where the number of correctly classified examples $c = 1, \dots, N$ and p is usually set to 0.5 in a two-class paradigm[5]. An example for the application of such Binomial models is given at a later point for illustrating the multiple comparisons problem in Section 5.4.

³since the values that X can take are discrete, a probability *mass* function is associated (instead of a *density* function as in the other examples of parametric tests)

5.2.4 Tests for normality

All above parametric tests assume the observations x_1, \dots, x_N to be drawn from a normal distribution. In order to test whether this assumption holds given experimental data, a *normality* test procedure can be implemented.

For determining, whether the underlying observations are normally distributed or not, I will use Shapiro-Wilk's W-test for normality[76], which had been found to be the most powerful test for normality[77], in particular for small sample sizes.

5.3 Nonparametric statistical inference

In certain cases the assumptions required for parametric tests are not met by the experimental data, such as if the data, for instance, is not distributed normally. The parametric assumptions are often not met the context of statistics on classification-derived data. A possible remedy of such situations is the recourse to *nonparametric* statistics. The most widely used nonparametric tests used in this context are resampling techniques, in particular permutation tests and bootstrapping methods.

5.3.1 Permutation tests

Permutation methods empirically construct a null distribution under the null hypothesis

H_0 : There exists no dependency between the class label and the data points

In other words, the null hypothesis states that manipulating the relationship between the labels and the data points does not have any impact on the results. The null distribution is derived by exchanging the class labels of the observations (e.g. by applying a random permutation to the order of observations while keeping the labels fixed). Next, a test statistic is computed using the new random relationship between class labels and data points. The procedure is repeated for many times, resulting in an empirical null distribution. This null distribution can then be used to assess the probability of the original (non-permuted) result in the light of the null hypothesis.

As an example, consider the thought experiment of Section 5.1 on page 31, where two groups each of ten patients suffering the same disease are given a new drug and a placebo medication, respectively. Let us assume that after a month of treatment the following health values are measured (by virtue of an idealized measurement procedure, which maps a value of 100 to perfect health and a value of 1 to near death):

Health given new drug: {73, 66, 63, 74, 75, 58, 67, 75, 72, 73}

Health given placebo: {69, 54, 69, 53, 60, 69, 63, 50, 70, 66}

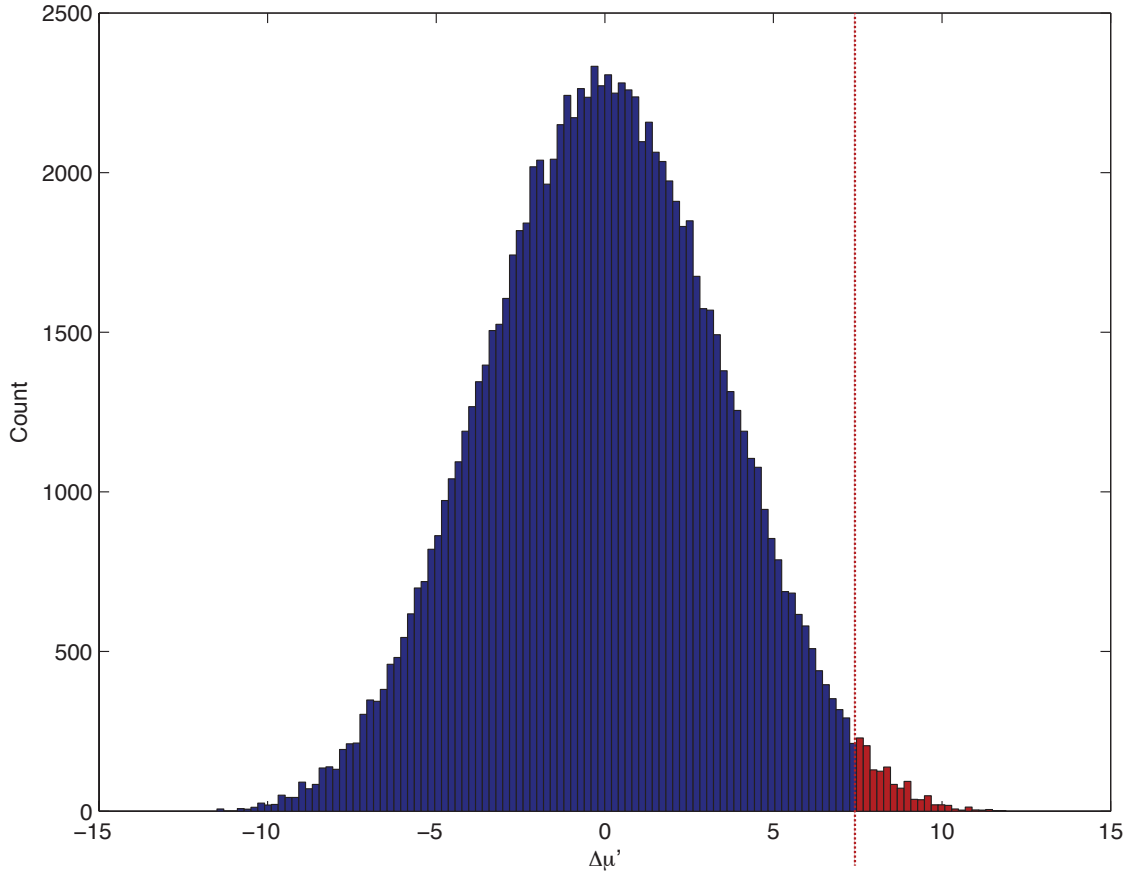


Figure 5.3: Histogram of the difference of health values *under permutation*: For 100000 repetitions, the participants were randomly distributed into the patient and control group and the difference between the mean health values of each group was computed. This resulted in an empirical null distribution. The original (non-permuted) difference of group mean $\Delta\mu = 7.3$ is marked as dotted red line. The probability of finding the original difference $\Delta\mu$ is the relationship between the right-tailed area (the red bars) and the rest of the distribution (the blue bars). As the red bars cover 1.31% of the total area of the histogram, the p -value for $\Delta\mu$ is 0.0131

The mean health of the new drug group is $\mu_{drug} = 69.6$, the mean health of the placebo group is $\mu_{plac} = 62.3$. The difference between both means is $\Delta\mu = \mu_{drug} - \mu_{plac} = 7.3$. Permutation methods allow the computation of the probability that the new drug actually has no curative effect, i.e. that it has the same effect as a placebo treatment. Hence, the probability of the original difference $\Delta\mu$ given chance is determined. Following the rationale of the permutation test, a random permutation to the order of the data points is applied, i.e. the observations are randomly shuffled and reassigned into both groups (allowing a *crossing* between the groups). For instance, one instance of shuffled data points may look like

Health given new drug: {50, 58, 66, 53, 54, 63, 69, 73, 75, 75}

Health given placebo: {66, 69, 72, 69, 70, 74, 63, 73, 60, 67}

Under the permutation, the difference between both means is $\Delta\mu' = \mu'_{drug} - \mu'_{plac} = 63.6 - 68.3 = -4.7$. The permutation procedure is repeated over many repetitions (e.g. 100000 times) and the difference between both means under permutation $\Delta\mu'$ is noted. Next, the histogram $H_{\Delta\mu'}$ of $\Delta\mu'$ is computed (as shown in Figure 5.3).

Using the null histogram $H_{\Delta\mu'}$, it is possible to compute the probability of the original difference (without any permutation applied) $\Delta\mu$:

$$p(\Delta\mu) = \frac{1}{N} \sum_{\Delta\mu' > \Delta\mu}^{\infty} H_{\Delta\mu'}$$

In here, the factor $1/N$ serves as normalization constant, rendering the overall sum of the histogram equal to 1. In our example the probability for the null hypothesis being true is derived as $p(7.3) = 0.0131$. As this probability is lower than the significance level $\alpha = 0.05$, it can be stated that the treatment using the new drug has a significant effect for curing the illness.

In summary, permutation methods aim to empirically construct the null distribution and use this null distribution for estimating the probability of the non-permuted result (under the null hypothesis). If the probability is lower than the significance level α , the null hypothesis is rejected.

Permutation methods have the property of an *exact* test, i.e. the probability of occurrence of a type I error is equal to α [74, page 39]. The permutation method relies on the assumption of *exchangeability* of the observations. For this, the joint distribution of the observations does not depend on the *order* of the subscripts of the observations[74, page 269]. In the illustrative example of patients this assumption is justified, as it appears unlikely that there exists a dependency structure between the health values of different patients.

5.3.2 Bootstrapping methods

The idea behind bootstrap methods is to resample repeatedly, however *with* replacement, from the original sample. For this, each group is resampled entirely from the samples within the group (with replacement) and the desired estimate is computed[78, page 8]. This allows to an empirical approximation of the (unknown) population distribution on the basis of the measured samples.

For illustration, suppose the same example as used above in the permutation statistics (see Section 5.3.1 on page 37), with health values of two groups of patients, one treated with a illness-curing drug and the other one with a placebo.

Health given new drug: {73, 66, 63, 74, 75, 58, 67, 75, 72, 73}

Health given placebo: {69, 54, 69, 53, 60, 69, 63, 50, 70, 66}

For each bootstrap draw, the health values of both groups are resampled with replacement, however for each group separately. For instance, this may results in

Health given new drug: {63, 67, 67, 72, 72, 73, 73, 74, 75, 75}

Health given placebo: {53, 54, 54, 60, 60, 63, 66, 69, 70, 70}

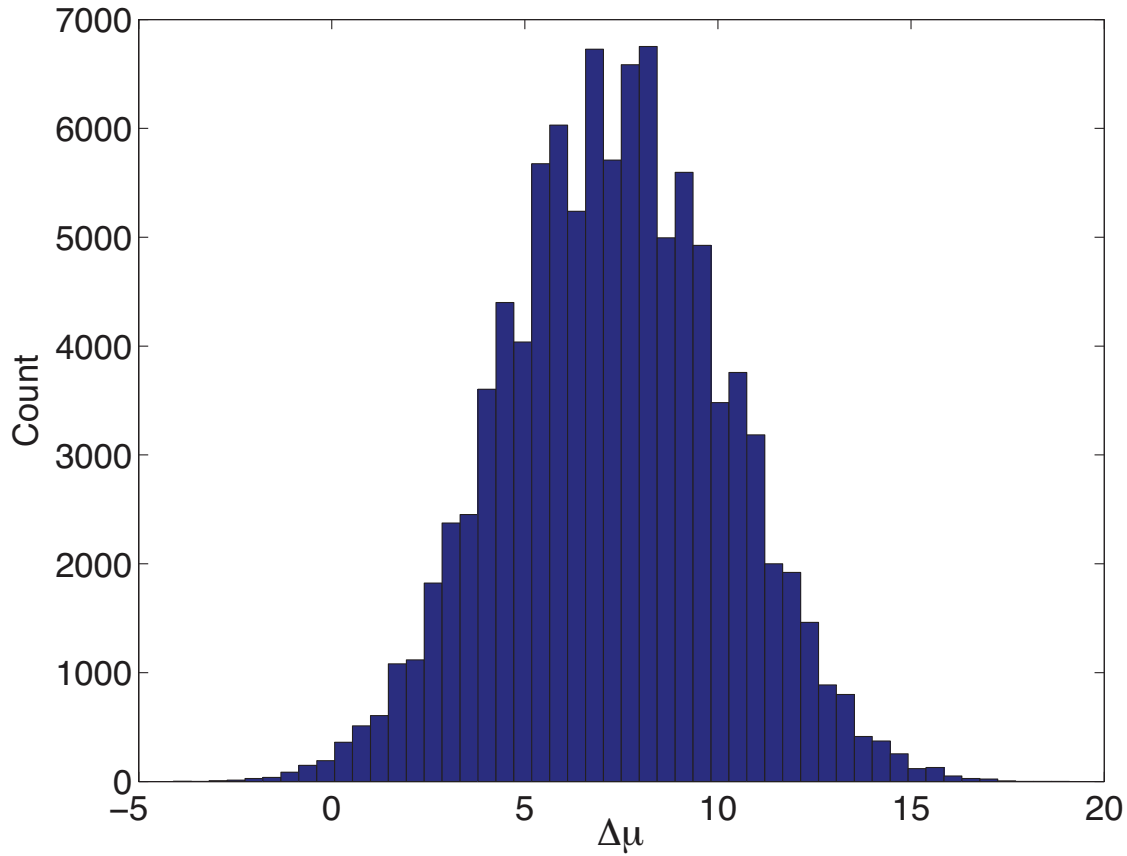


Figure 5.4: Approximation of the population of exemplary health data *using the bootstrap method* and $N = 100000$ resampling steps.

The difference between both means in this example is $\Delta\mu' = \mu'_{drug} - \mu'_{plac} = 71.1 - 61.9 = 9.2$. The procedure is repeated for N repetitions (e.g. $N = 100000$), each time the difference $\Delta\mu'$ is noted, resulting in an *approximation* of the population distribution. The approximation of this example is displayed in Figure 5.4.

It should be noted that the application of the bootstrap procedure for this illustrative example may be ill-advised, as bootstrap procedures are generally not recommended for sample sizes of less than 100 observations[78, page 19].

In my work a similar technique is used for computing group statistics. As the procedure is not adequately classified as a *bootstrap* procedure, I will refer to it as *Monte-Carlo resampling*.


5.4 The multiple comparisons problem

In all above examples for statistical testing I have described significance testing for *one single* statistical test. However, if a set of *many* statistical tests is carried out *simultaneously*, it is possible that some number of these tests return significant results only due to the fact that *many* tests had been carried out.

As an example, a simple dice experiment can be considered, using a standard dice

which has six faces (which conveniently are numbered from 1 to 6). The assumption that the dice is fair can be tested, i.e. if each face is being equally likely with the probability of $p_i = \frac{1}{6}$ if thrown once. The experiment is comprised of three successive throws of the (same) dice. Since each of the successive three throws is *independent*, it is possible to use a binomial model to compute the probability that *all three* throws show each the number 6.

$$p_{666} = f(3, 3, \frac{1}{6}) = \binom{n}{k} p_6^k (1 - p_6)^{n-k} = \left(\frac{1}{6}\right)^3 \approx 0.0046$$


In other words, the probability p_{666} of three successive trials that each show face number 6 (in the following abbreviated by ) is *very* low and around 0.46%.


Let us put forward the null hypothesis for a statistical test:

H_0 : the dice is fair

and the complementary alternative hypothesis

H_1 : the dice is unfair⁴ and biased towards showing the face with number 6.

Using a significance level of $\alpha = 0.05$, the null hypothesis H_0 (of a fair dice) would be rejected in case the outcome of the experiment were , as the probability for this is very low given a fair dice and below the significance level, i.e. $p_{666} \ll \alpha$.

However, if one would use this experimental procedure for testing the fairness of more than one dice, a problem known as the *multiple comparisons problem* would arise: The probability that *at least one* of the multiple dices would show  is *much* larger than the previously calculated probability p_{666} . In other words, our single-dice experimental procedure and statistics are inappropriate if we test many dices, as even a fair dice can come up three times successively showing the number 6 if many dices are thrown. Let us for now assume that we want to test the fairness of $N = 10000$ dices with the above method. Analytically, it is possible to derive the probability that *at least one* of the dices will show a pattern of three successive trials with face 6, given that all dices are *fair*:

$$p_{pos} = 1 - p_{none}$$

where p_{pos} is the probability that at least one of the dices shows three successive trials showing face number 6, and p_{none} is the probability that *none* of the dices shows three successive trials with number 6. Using another binomial model, it can be derived that

$$p_{none} = (1 - p_{666})^{10000} \approx 7 \cdot 10^{-21}$$

In other words it is extremely unlikely, that *not a single one* of the 10000 fair dices does show a pattern of three successive trials. Hence, in good approximation, $p_{pos} \approx 1$, i.e. it is practically almost certain that at least one of the fair dices shows three successive trials with face 6.

⁴consider for instance a dice where the center of gravity is not at the geometric center, thus introducing a bias

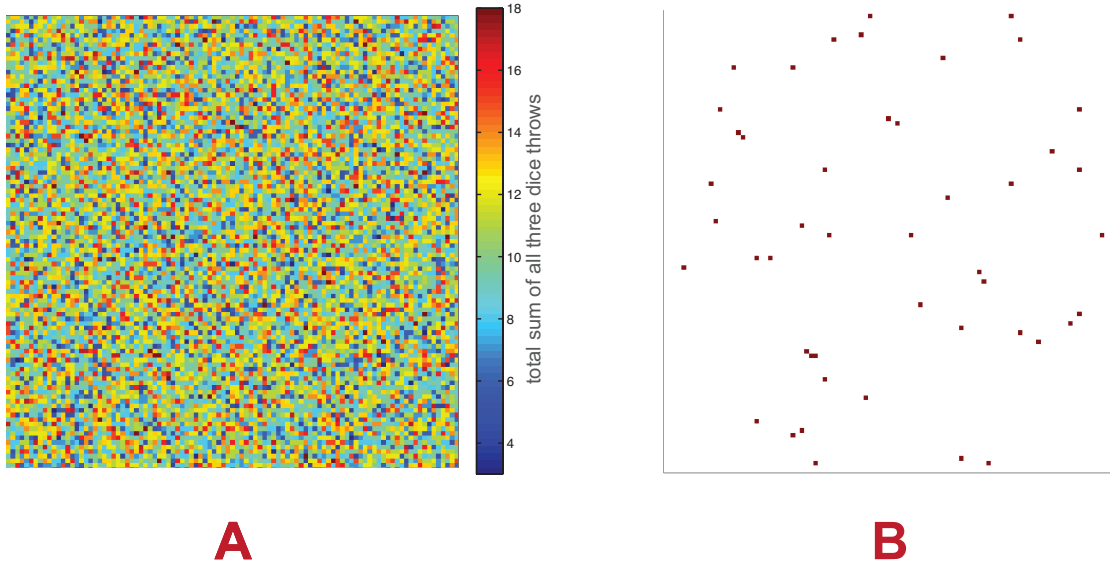


Figure 5.5: Visualization of the multiple comparisons problem, which implies that even rare events can occur by chance if there are enough repetitions. **(A)** Sum of three subsequent dice throws, repeated for 10000 times. For better visualization, all trials have been rearranged in a 100x100 matrix. The bluest color, which stands for a sum of 3, implies the dice throw combination 1-1-1. The reddest color corresponds the sum 18 which only is achieved for the combination 6-6-6 **(B)** The matrix is displayed in thresholded form, where only the sum 18 is above the threshold. Hence all red dots imply a dice throw combination 6-6-6. In total, there were 51 trials with this combination. In other words, although the probability of three successive trials for a single dice throw experiment with the outcome of three times face 6 was thrown was very low (about 0.46%), if repeated often enough it is exceedingly likely to happen.

For the purpose of illustration, I will provide a toy simulation of the above example: Three dice throws are simulated by randomly selecting three numbers between 1 and 6 with uniform probability; hence the virtual dice throws are *a priori* fair. The procedure was repeated for 10000 times. For visualization, only the sum of the three trials is taken into account and the data is rearranged in a matrix of size 100 x 100 and displayed in Figure 5.5A. Hence in this figure, a value of 18 corresponds to three successive trials where number 6 was selected. In Figure 5.5B, a threshold was applied, so that only number 18 is displayed. In total, the combination $\{\{6\}\{6\}\{6\}\}$ occurred 51 times in this simulation, which (in good approximation) corresponds to the product of the probability p_{666} of three successive trials where face 6 is up with the number N of repetitions⁵.

Conclusively, statistical testing procedures that imply many tests may result in exceeding rates of false positivity if no *correction* for the multiple testing is applied. This is especially critical for fMRI data, where the number of statistical tests commonly is equal to the number of voxels, in other words extremely large. In the following, I will discuss the most important multiple comparison correction methods currently used in fMRI.

⁵if the simulation was repeated often, the mean number of 6-6-6 occurrences would converge to $N \cdot p_{666}$

5.4.1 Bonferroni correction

The Bonferroni correction is the most stringent and conservative multiple-testing correction. It states that if N statistical tests are carried out simultaneously, the significance level α has to be adjusted to [74, page 79]

$$\alpha_{bonf} = \frac{\alpha}{N}$$

A Bonferroni-corrected test has the property of *strong control*, as the number M_{fp} of false positives (where the null hypothesis is rejected but in truth holds) is constrained by α_{bonf} [79]:

$$E\left(\frac{M_{fp}}{N}\right) \leq \alpha_{bonf}$$

where $E(\frac{M_{fp}}{N})$ is the expectation value for the rate of false positivity.

However in case of fMRI data, where N typically is very large ($N \geq 50000$), the corrected significance level α_{bonf} is overly stringent. Effectively, whole-brain statistics using Bonferroni type I error correction would label practically no voxels significant. On the other hand, the type II error rate (false negativity) is drastically increased, rendering the Bonferroni correction not sensitive enough to be useful for the statistical analysis of fMRI data [80].

The Bonferroni method implements the (pessimistic) assumption, that all carried out tests are *independent* from each other (however independency is not required). This assumption, however, is not realistic in case of fMRI data, since neighboring voxels feature a spatial correlation [27] (and hence are not fully independent from each other). In the following, I will briefly sketch out other strategies for the correction of the multiple testing problem that are tailored for dealing with fMRI data.

5.4.2 False discovery rate

The false discovery rate (FDR) is defined as the proportion of false positives among the subset of tests, where the null hypothesis was rejected. This is in contrast to the Bonferroni method described above, as the Bonferroni method controls for the false positives over *all* tests performed, regardless whether the null hypothesis was rejected or not [80]. In other words, FDR methods control the false positivity rate *exclusively* for the tests that have been labeled *significant*. Given M_{sign} tests that are labeled as significant, and among these M_{fp} tests are erroneously labeled significant (i.e. total number of false positives), the *rate* Q of false discovery is defined as [81]:

$$Q = \frac{M_{fp}}{M_{sign}}$$

Procedures controlling the false discovery rate then ensure the following:

$$E(Q) \leq q$$

Conventionally, q is set between 0.01 and 0.05. In the original formulation of the FDR

procedure[81], the set of p -values of each of the N hypotheses being tested are rearranged in ascending order:

$$p_1 \leq p_2 \leq \dots \leq p_N$$

corresponding to the hypotheses H_1, H_2, \dots, H_N , which were reordered according to their respective p -values. Then the following *cutoff criterion* is proposed: let k be the largest index i for which $p_i \leq \frac{i}{N} \cdot q$ holds, then reject the hypotheses H_1, \dots, H_k . Note that given the case that the above condition does not hold for *any value* of k , *no* hypotheses are rejected.

In case of a large number N of hypotheses being tested and a *correlation* amongst the hypotheses, another formulation is used. Instead of defining the cutoff for p -values smaller than $\frac{i}{N} \cdot q$, another, less conservative cutoff condition can be employed: let k be the largest index i for which $p_i < \frac{i}{N} \cdot q \cdot \frac{1}{c(V)}$, where $c(V) = \sum_{i=1}^N \frac{1}{i}$ is the harmonic series. As before, the hypotheses H_1, \dots, H_k are rejected then. The preference for the latter cutoff criterion in case of fMRI data is motivated by the dependency between hypotheses, as there exists a spatial correlation across voxels[80].

Alternatively, in case of a small number N of hypotheses being tested, a step-down FDR procedure can be applied[82]. Firstly, N critical values are defined by:

$$\delta_i \equiv 1 - \left[1 - \min\left(1, \frac{N \cdot q}{N - i + 1}\right) \right]^{\frac{1}{N - i + 1}} \text{ for } i = 1, \dots, N$$

The cutoff criterion is defined as the following: let k be the largest i for which $p_i > \delta_i$. Then all hypotheses H_1, \dots, H_{k-1} are rejected.

5.4.3 Random field methods

The multiple comparisons methods described above apply for *voxel-wise* statistical inference, i.e. each voxel or location corresponds to a single hypotheses. Alternatively, methods tailored for the characteristics of fMRI data have been proposed, which are not applied on the voxel-level but rather establish inference on *topological features*[83], which consists of many voxels. This greatly reduces the amount of hypotheses tested and hence alleviates the multiple comparisons problem. Commonly, topological features are defined as clusters (i.e. connected excursion sets) of voxels, where each voxel exceeds some predefined probabilistic threshold.

Gaussian random field methods are one class of topological inference methods. The underlying idea behind this multiple comparisons strategy is to assign a probability (i.e. a p -value) to a cluster given the observed attributes[83](e.g. image smoothness, cluster size etc.). If the probability for a cluster is small and below the predefined cluster-level α , the cluster is labeled significant. The derivation of the probability of finding a supra-threshold cluster is based on a parametric distribution approximation of cluster attributes, requiring several assumptions on the data[84]. Most importantly, the data is assumed to be smooth (i.e. there exists a high spatial correlation between neighboring voxels) and the smoothness of the data is uniform across the brain. This requires spatial smoothing of the data, which commonly

is performed by a Gaussian smoothing kernel with a full width at half maximum (FWHM) between 3 and 8mm.

In the framework of the Gaussian random field theory, fMRI data is considered as a lattice approximation of a smooth Gaussian random field[85, 86, 87] and furthermore sufficiently high voxel-wise thresholds p_{vox} (to be included into a cluster) are required. As a full derivation of the cluster statistics given random field theory is out of the scope of my thesis, I will only provide the most important results: the approximation for the distribution for cluster size s' of a three-dimensional Gaussian random field, which is used in the SPM software package[29], is derived as[84]:

$$Pr\{s' > s\} \approx e^{-\psi s^{2/3}} \quad (5.8)$$

where $\psi = \left[\frac{\Gamma(5/2)E(L)}{E(U)} \right]^{2/3}$ and the expectation value of the search volume L is defined as $E(L) = \sum_{d=0}^3 R_d \rho_d(p_{vox})$. This implements the *resel* count R_d , which is the volume measured in terms of the *estimated* image smoothness and their respective densities ρ_d . The expectation value for the distribution of the supra-threshold volume[84] is defined as $E(U) = V(1 - F(p_{vox}))$, where V is the overall search volume and F the underlying cumulative distribution function of a Gaussian random variable.

5.4.4 Non-parametric cluster statistics

Permutation based methods for solving the multiple comparisons problem have been pioneered in the mid 90's[79] and became increasingly popular in the following years[88, 2, 84]. The rationale behind these methods is to construct a pool of *chance* images (maps) from the underlying data using a large number of permutations on the data labels. For example, the data may consist of anatomical images of two subject groups (patients and control). A plausible null hypothesis here would state that there is no difference between patients and control in terms of the recorded images.

Following the rationale of the permutation method (as already described in Section 5.3.1 on page 37), a random permutation is applied to the labels of the data points (patient and control). Crucially, this permutation is applied to the *whole image* at once (and not voxel-wise), yielding one chance map per permutation. On the chance maps, the desired test statistic is computed (e.g. T-tests). Repeating this procedure many times yields a *pool* of chance statistical maps, which are thresholded voxel-wise using the threshold p_{vox} for the given test statistic. Next, a search for supra-threshold clusters is performed in this pool, yielding a cluster size distribution under the null hypothesis. The same cluster search is performed in the test statistic given the original (non-permuted) images. Ultimately, using the chance cluster distribution, each cluster size of the original data can be assigned a probability. As each assignment between cluster size and cluster probability represents a statistical test, a multiple comparisons correction on cluster level has to be applied (commonly a Bonferroni correction or FDR methods are used for this).

Part II

Methods

Chapter 6

fMRI data sets

In total, two different fMRI data sets were used for my thesis. The data sets were recorded in different resolutions, one with a rather standard resolution and the second with an ultra-high resolution (see Figure 6.1). Both studies were finger tapping studies, however they comprised different experimental paradigms.

The first study had been performed with a 3T system and whole-brain coverage (*3T tapping synchronization experiment*). The aim of the study was to measure the subject's brain response while they were synchronizing to auditory or visual pacing sequence, which was either continuous or discrete. Hence in total, the experiment consisted of four conditions. The data set was originally recorded by Dr. Mike Hove.

The second study was acquired on a 7T system (*7T finger tapping and imagination*) and only covered a small region in the left hemisphere of the subjects (see Figure 6.2). The paradigm consisted of four conditions; one rest condition, where subjects were not instructed with any task, and two finger movement conditions, where subjects either tapped freely with four fingers or tapped with four fingers touching their thumb. In the last condition, the subjects were instructed to imagine tapping of the four fingers without moving them. The data set was originally recorded by Dr. Robert Trampel.

6.1 3T tapping synchronization experiment

6.1.1 Experimental design

Fourteen healthy subjects participated in the study[89, 5] (age range 24 to 34, all right-handed). The subjects were paid for their participation and gave written consent. The data sets of two subjects were incomplete and therefore were excluded from further analysis, leaving a total of 12 participants. The experiment consisted of 80 trials, which were split into two recording sessions. Each trial lasted 19.2s and was separated by a variable inter-trial interval of between 9.4 and 12.2s. Participants were instructed to tap with their right index finger in

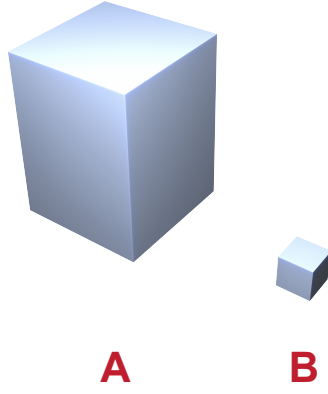


Figure 6.1: Illustration showing the difference in resolution of the two fMRI experiments used in this thesis. **(A)** 3T tapping synchronization experiment, voxel size = $3 \times 3 \times 4 \text{ mm}^3$, $B_0 = 3T$ **(B)** 7T finger tapping and imagination, voxel size = $0.75 \times 0.75 \times 0.75 \text{ mm}^3$, $B_0 = 7T$

time with four isochronous experimental conditions: a discrete auditory pacing sequence (50ms sine beeps at 1350Hz), a continuous auditory pacing sequence (pitch sweeps between 1350Hz and 450Hz, $T_{\text{cycle}} = 600\text{ms}$), a discrete visual sequence (a white bar flashed for 50ms over a black background), and a continuous visual sequence (a white bar moving up and down). In the fMRI study, each of the four conditions was presented with a slow and a fast variant (inter-stimulus-interval $\Delta t_{\text{event}} = 400\text{ms}$ or $\Delta t_{\text{event}} = 600\text{ms}$). Throughout this thesis, the slow and fast variants of each condition were merged together. The presentation of the pacing sequences was randomized using a computer and the software Presentation[90], which also recorded the tap timing[5].

For the sake of simplicity and clarity, only two the two visual experimental conditions are analyzed throughout this thesis, the two auditory conditions were not used.

6.1.2 Data acquisition

Functional MRI data (gradient EPI) was collected on a 3T system (TRIO 3T, Siemens Healthcare, Erlangen, Germany) with a standard head coil[5]. The scans contained 36 axial slices covering the whole brain ($TR = 2000\text{ms}$, $TE = 24\text{ms}$, slice thickness 4mm with 1mm gap, in plane resolution $3 \times 3 \text{ mm}^2$). A sagittal T_1 -weighted anatomical scan was obtained from the database of the Max-Planck-Institute for human cognitive and brain sciences for all subjects (3T Siemens Trio system, $TR = 1300\text{ms}$, $TE = 3.93\text{ms}$ and an isotropic voxel size of 1 mm^3).

6.1.3 Data preprocessing

The data was corrected for head motion (as described in Section 8.1 on page 61), coregistered to the anatomical scan (using a rigid body transformation with three spatial and three rotational degrees of freedom) and spatially normalized to the MNI305 space (see 8.3). A temporal high-pass filter with a cutoff-frequency of $f_{\text{highpass}} = \frac{1}{80}\text{s}$ was applied to the data to remove low frequency drifts (as described in Section 8.2 on page 61). After this, a standard GLM was

fitted to each experimental trial to estimate its β -parameters (see Section 8.4 on page 63). Hence, for each of the four conditions 20 three-dimensional β -maps were obtained. However, as stated before, the two auditory conditions were not further regarded in this thesis; only the two visual conditions are analyzed.

6.2 7T finger tapping and imagination

6.2.1 Experimental design

Ten healthy subjects took part in the study (age range 23 to 28, all right-handed). All subjects were paid for their participation and gave written consent. The data sets of two subjects were incomplete and were discarded from further analysis, leaving a total of 8 participants. As the normalization to the standard MNI305 space was not feasible for data covering only a part of the brain with a very high resolution, one single subject was selected¹, omitting group analysis. The experiment was comprised of 15 trials per experimental condition, each lasting for 26.4 seconds. In total, the experiment consisted of four experimental conditions, which were presented subsequently (without randomization) in a block-design fashion: The first condition was a rest condition, where subjects were instructed not to move and not to imagine movement. Following the rest condition the subjects were asked to imagine finger tapping, without involving actual finger movement. This was followed by the tapping condition, where subjects sequentially tapped with the four fingers of their right hand to the thumb of the right hand. No external pacing sequence was used, however, subjects were asked to maintain a frequency of $2Hz$. In the last condition, subjects performed the same tapping sequence as before, however, without touching their thumb.

In this study, two conditions (rest and tapping with touch) were selected for further analysis, the other two remaining conditions were not included here.

6.2.2 Data acquisition

The experiment was performed using a 7T system (MAGNETOM 7T, Siemens Healthcare, Erlangen, Germany), using a 24 channel head coil (NOVA Medical Inc., Wilmington MA, USA). T_1 -weighted structural scans were acquired using a MP2RAGE[91] scanning sequence ($TR = 8250ms$, $TE = 2.59ms$ and a isotropic voxel size of $(0.7mm)^3$). The functional scans contained 17 to 31 axial slices (depending on the subject) covering the left motor cortex of the subject's brains ($TR = 3300ms$, $TE = 25ms$, slice thickness $0.75mm$, in plane resolution $0.75 \times 0.75mm^2$) using a novel acceleration technique[92].

¹a representative subject was selected. All results given in this thesis are comparable throughout all subjects of this experiment

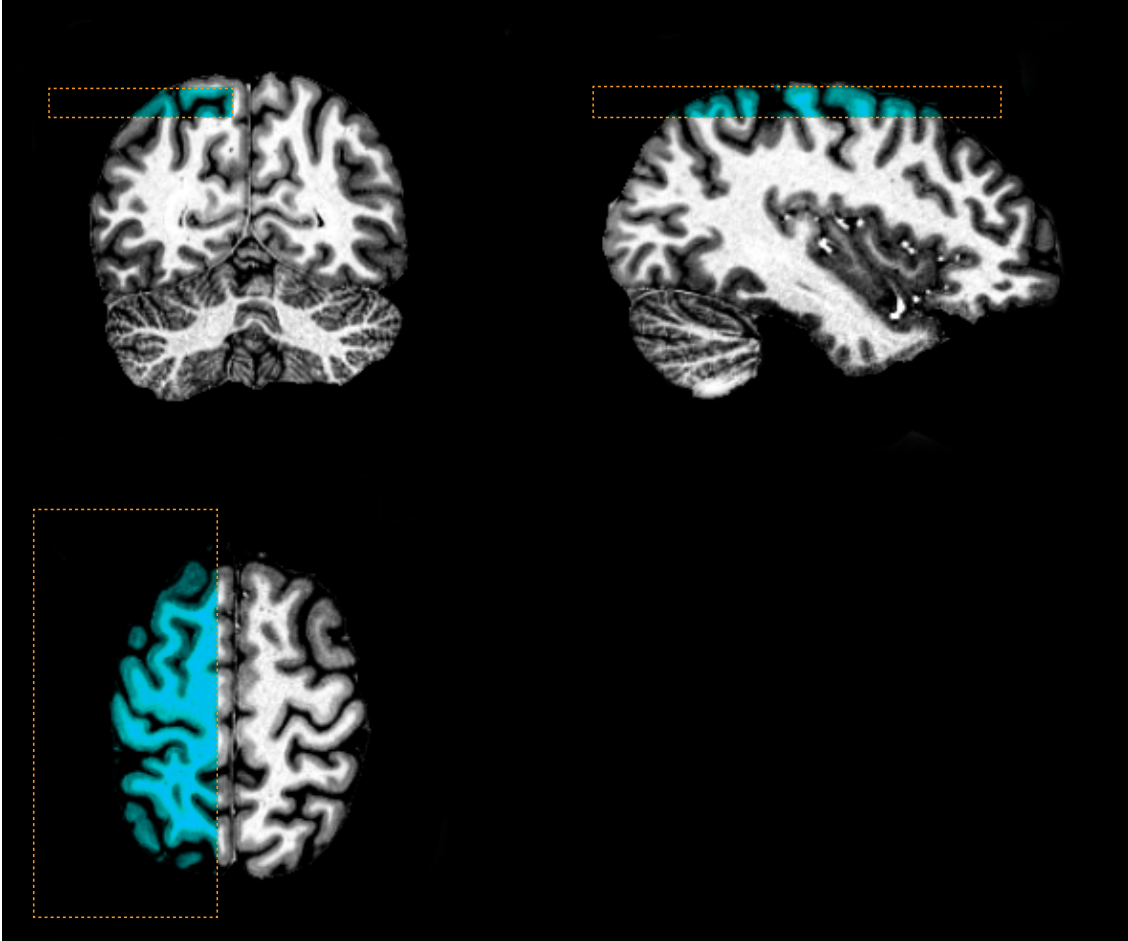


Figure 6.2: Brain coverage of the ultra-high resolution fMRI data set. Only a small part of the left hemisphere (marked in blue color) containing motor, sensory, parietal and frontal regions was scanned since a whole-brain coverage would result in too slow of a scanner repetition time TR at the isotropic resolution of 0.75mm.

6.2.3 Data preprocessing

Head motion correction was carried out as first preprocessing step using SPM8[29] (see Section 8.1). Low frequency drifts were removed using a temporal high-pass filter with a cutoff-frequency of $f_{highpass} = \frac{1}{80}s$ (see Section 8.2). Subsequently, a standard GLM was fitted to each experimental trial to estimate its β -parameters (see 8.4). This resulted in a total of 15 three-dimensional β -maps per experimental condition. As well as in the other fMRI study, only two conditions were selected for further analysis, namely the rest and tap condition (with touch). The two remaining conditions were not further analyzed in this work.

Chapter 7

Synthetic data sets (simulations)

7.1 Single Subject Simulations

The goal of the single subject simulations was to artificially recreate volumes with pre-defined properties and to analyze these with identical methods as the fMRI data. In particular, the simulations allowed the manipulation of the geometry of the distribution of information in a meaningful way, mimicking the geometry of cortical activation patterns in a controlled fashion. This allowed a quantitative investigation of the behavior of the two information mapping methods used in this thesis (the feature weight mapping method and the searchlight decoding method). Furthermore, I created a null simulation to validate the false positivity rate.

The nonparametric statistics of all simulated single subject data sets was carried out on the permutations directly (without the Monte-Carlo resampling procedure). For this reason, in 10000 permutations were computed for each data set except for the null simulation, where for computational reasons only 1000 permutations were obtained.

7.1.1 Single subject geometric simulation

A data set representing *one virtual* subject was created. The dataset was comprised of 30 volumes in total; 15 volumes were assigned to class *A* and 15 to class *B*. Each volume had the size $66 \times 22 \times 22$ voxels. The volumes of both classes were filled with Gaussian noise of the normal distribution $\mathcal{N}(0, 1)$ and smoothed slightly with a Gaussian smoothing kernel, the FWHM of the kernel was set to one voxel. After this, an offset of size 0.5 was added at three locations in class *A* and three different locations in class *B*. These locations constituted three half-cubes for each class, positioned at the centerline of the volume (see Figure 7.1). For class *A*, the offset was added for the three upper half-cubes, for class *B* the offset was added for the three lower half-cubes. Importantly, the size of the half-cubes varied; the leftmost half-cube was of the size of $6 \times 6 \times 1$ voxels (representing *fine* information spread), the second one had the dimension of $6 \times 6 \times 2$ voxels (representing *medium* information spread) and the rightmost half-cube $6 \times 6 \times 3$ voxels (*coarse* information spread). Effectively there was a 4-voxel gap



Figure 7.1: Information spread of the single-subject geometric simulation. Information was deposited in a total of 6 half-cubes. For class *A*, information was deposited in the three upper half-cubes, displayed in violet color, while for the three blue half-cubes information was present only for class *B*. The half-cubes varied in their size in *z*-direction and hence the size of the gap between the half-cubes.

between the leftmost, a 2 voxel gap between the middle, and no gap between the two rightmost half-cubes.

7.1.2 Single subject null simulation

100 single-subject null data sets were generated. The intention of the simulation was to empirically validate the false-positivity rate, hence there was no class information deposited at any location. The data sets constituted each 30 volumes (15 for class *A* and 15 for class *B*). The volumes were blocks of the size $40 \times 40 \times 40$ voxels and were filled with noise drawn from a uniform distribution in the interval $[0, 1]$. The volumes were smoothed with a Gaussian smoothing kernel, the FWHM of the kernel was set to one voxel. Next, an area of $30 \times 30 \times 30$ voxels was cut out from the center of the volume in order to avoid distortional effects at the border of the volumes due to the spatial smoothing. For computational reasons, only 1000 permutations were carried out for the SLD approach, the number of permutations for the FWM approach was set to 10000.

Furthermore, I created 10 single-subject simulations as above, however with differing levels of smoothness (by varying FWHM of Gaussian smoothing kernel). For this I used 10 equidistant values between 0 and 3 voxels. This procedure allowed the empirical investigation of the impact of the intrinsic smoothness on the nonparametric framework.

7.2 Group Simulations

The point of the group-level simulations was to emulate group data sets with pre-defined properties and to analyze them with the different statistical frameworks. On one hand this allowed a detailed comparison between parametric T-based frameworks with the proposed nonparametric framework; on the other, the two information mapping methods (FWM and SLD) could also be compared between each other.

7.2.1 Group simulation 5 cubes

Twelve data sets representing *virtual* single subjects were created[5], and all subsequent analysis was carried out identically as the fMRI data sets. Each simulated data set consisted of 16



Figure 7.2: Information spread for the group simulation 5 cubes. The offset which was added to the background noise depended on the cubes location, the leftmost one had the smallest amount of information deposited while the rightmost one had the highest information content.

volumes in two classes (eight for A and eight for B). The volumes comprised blocks of the size of $108 \times 17 \times 17$ voxels, the format was chosen for illustrative reasons. For class A , the volumes were filled with uniformly distributed random numbers sampled from the interval $[0, 1]$. The volumes for class B were created in the same way, but at specific spatial locations an offset was added. Therefore, only in these locations was information about the class present. The spatial locations where an offset was added were five *cubes* with an edge length of six voxels, which were aligned in the middle of the volumes with equal distances. Within the cubes, I added an offset between 0.15 and 0.2 to the uniform noise (with the most left cube having an offset of 0.15 and the most right at 0.2, see Figure 7.2)[5].

Furthermore, two *information degradation* procedures were applied[5], since in real fMRI group data there are two distinct sources of variability: On one hand, the neural activity may be dependent over time, i.e. the response elicited by the same stimulus presented at different times may systematically vary with time. This source of variability is known as *inter-session* or *inter-run* variability. The second source of variability stems from the fact, that neuronal response profiles depend on the subject (e.g. differences in individual anatomy or cognitive strategies). Consequently, this source of variability is known as *inter-subject* variability. In order to account for the first source of variability, the inter-run variability, a random value between 0% and 50% of the offset of the corresponding cube was subtracted for each of the eight volumes (in class B). To account for the inter-subject variability, I randomly subtracted 0–50% of the offset of the corresponding cube for each data set, in other words the same percentage was subtracted for all eight volumes. Therefore, the information content in the cubes depended on the session, the virtual subject and the position of the cube, with the leftmost cube having an offset between 0 and 0.15 and the rightmost cube having an offset between 0 and 0.20[5].

7.2.2 Group null simulation

In total, 100 group null data sets were generated [5]. No class information was deposited at any location, as the intention of this simulation was to empirically validate the false-positive rate. Each of these group data sets consisted of 10 *virtual* single-subject data sets. Each of the single-subject data sets consisted of 10 volumes (five for class A and another five for class B). The volumes comprised blocks of $30 \times 30 \times 30$ voxels. All volumes were filled with noise drawn from a uniform distribution in the interval $[0, 1]$.

7.3 General simulations

Besides single-subject and group-level simulations, I created two further types of simulations. The first simulation investigates the influence of the applied cross-validation scheme to the distribution of classification accuracies. This is especially relevant for the binomial models, which are assumed to not show any difference in regards to the underlying cross-validation scheme. The second simulation investigates empirically how many permutations are necessary on single-subject level for the Monte-Carlo group recombination procedure.

7.3.1 Cross-validation influence simulation

In order to test the effect of correlation between cross-validation folds, six different scenarios were created. Each scenario was repeatedly computed for 10^6 repetitions and consisted of 100 observations per class. The data points were filled with uniform noise of the interval $[0, 1]$. The number of features was set to 10 (for computational reasons). The scenarios differed in the number of applied cross-validations: I applied 2, 5, 10, 20, and 50 cross-validations. The size of the respective test sets were as follows: 50, 20, 10, 5, and 2, so that the product between number of cross-validations and size of the test set was constant for all scenarios. In the sixth scenario the classifier was trained 10^6 times on 200 samples (100 observations per class) and then tested on another, *completely unseen* data set of 200 samples, *without* applying cross-validation. For each scenario and repetition, the sum of correctly identified labels over all cross-validation folds was noted.

To quantify the deviation between theoretic binomial distribution (200, 0.5) and the six empirical distributions, the following error term was used:

$$\sigma = \sqrt{\frac{\sum (H_{bin}(i) - H_{emp}(i))^2}{n}} \quad (7.1)$$

where $H_{bin}(i)$ is the i -th entry in the histogram of the binomial distribution and $H_{emp}(i)$ is the i -th entry of one of the six empirical histograms. The i -th entry of the histograms indicates how often a number of i labels were predicted correctly (or are expected to be predicted correctly in case of the binomial histogram).

7.3.2 Simulation undersampling the permutation space

On the group level it is a priori not clear to which number of permutations on the single-subject level are required. In order to investigate the influence of the number of single-subject permutations on the group result, I undersampled the overall available permutation space using different levels of undersampling.

For this I simulated datasets consisting of 12 virtual subjects and their respective data matrices Y_i , consisting of a fixed number of features (5 features) and 16 examples in total. Half

of the examples were assigned to class A , the other half to class B . The data matrices were filled with noise from a uniform distribution of the interval $[0, 1]$.

I varied the number of permutations on single-subject basis using four levels of under-sampling: either 10, 100, 1000 or 10000 permutations were carried out. After the single-subject permutations were obtained, a group statistic was computed using Monte-Carlo resampling (see [Section 9.5 on page 70](#)). This yielded one null distribution H_{emp} for each of the four levels. Next, a normal distribution was fitted to H_{emp} , using the *normfit* function built-in MATLAB. As a result, the two best fitting parameters μ and σ of the normal distribution were estimated. For each level of undersampling, the simulation was repeated 1000 times (including data generation at each level).

Chapter 8

Preprocessing of the fMRI data

8.1 Motion Correction

Although experimental subjects are carefully instructed to avoid movement while inside of the MRI scanner, it is virtually impossible to remain perfectly still for any living human being. Subject motion manifests itself as a complex function on the fMRI data [93], since the MRI signal depends not only on the current position but also on the spin excitation history. Furthermore, movement causes a spatial shift in the resulting image matrices.

To compensate for the head motion, the movement effects are corrected post-hoc by computational means. Most commonly, a rigid body transformation with 6 degrees of freedom is applied, where three degrees are translational and the other three rotational. This most simple type of movement correction based on realignment was applied for all fMRI experiments in my thesis, using an implementation of the fMRI analysis software package SPM8[29].

It should be mentioned that pattern based analysis methods, such as the ones described in my thesis, are especially prone to movement related problems: On one hand, no spatial smoothing is applied beforehand and on the other it is intuitively accessible that any classifier will perform worse if a spatial pattern is shifted in its coordinates for subsequent examples (i.e. scans). Furthermore, motion related problems scale with the spatial resolution, since also smaller magnitudes of movement may have an effect in high-resolution scanning, while the same movement does not have a considerable effect for low-resolution data sets (see Figure 8.1).

8.2 Temporal filtering

The fMRI signal underlies fluctuations on all timescales and can be considered non-stationary. The fluctuations have a variety of sources such as low-frequency drifts caused by the scanner[94], and also importantly, spontaneous neuronal activity. The latter partly is due to the *intrinsic nature* of the brain as a system [95] and may reflect ongoing large-scale interactions and adaptations between different neuronal networks. Furthermore, there are physiological effects due

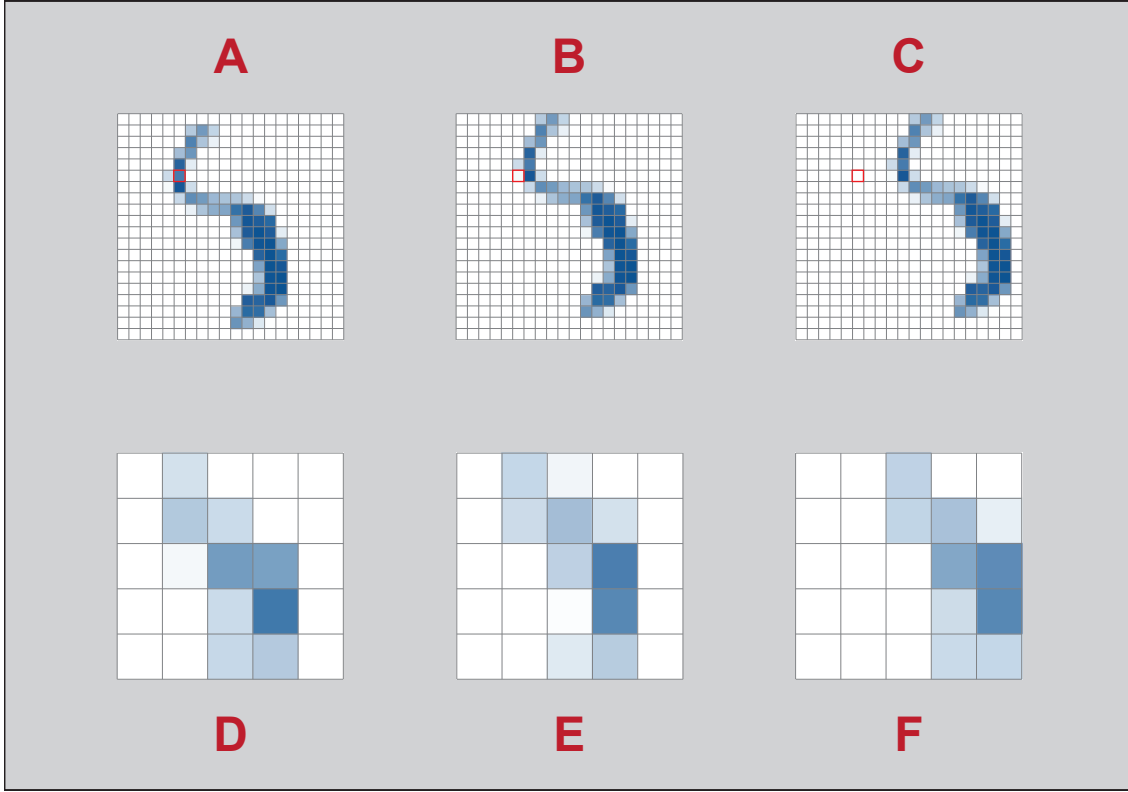


Figure 8.1: Effects of subject head motion on a high-resolution data set (upper row, simulating a voxel size of 0.75mm) and a data set in lower resolution (lower row, simulating a voxel size of 3mm). The columns represent time steps; the displacement between the time steps has the same physical size regardless of the resolution of the data sets. The small red square in the upper row is a highlighting of one voxel for illustrative reasons. The blue voxels represent activated areas. **(A)** 1st time step in the high-resolution data set. **(B)** 2nd time step introducing a small movement in the high-resolution data. It is clearly visible that the entire pattern has shifted, a large fraction of the pixel values has changed (e.g. the red square). **(C)** 3rd time step with a heavier movement in the high-resolution data set. The values for the largest part of the voxels has changed. **(D)** 1st time step in the low-resolution data set. **(E)** 2nd time step in the low resolution data set, introducing a small translation of the same physical dimension as in *B*. Most voxels have a comparable value as in the 1st time step of the low-resolution data set. **(F)** 3rd time step of the low-resolution data set, introducing a larger movement. The large movement has a high impact on the data set, about half of the (activated) voxels change their value to a considerable degree.

to cardiac and respiration cycles and changes respiratory flow rates[94].

Since the underlying idea of the pattern analysis methods discussed in this thesis is to map experimental control variables onto statistical changes in brain signals (see Section 4.1 on page 17), only signals on the temporal scale of the experimental trials are of interest. Ultimately, the low frequency content is not considered here and filtered out using a high-pass filter with a cutoff frequency of $f_{hp} = \frac{1}{80} s$.

8.3 Normalization to standard brain space

Human brains differ from person to person. While not only shape and volume are varying, *relative* positions of anatomical landmarks are also diverse. On the other hand, it is often desired to analyze data on a *group-level*, as only the most consistent patterns of activation related to a cognitive function are of interest. Hence, a spatial transformation of the data into a *common* coordinate space is necessary. This transformation is also known as *spatial normalization*. Most commonly, each brain is transformed onto an anatomical template given by the Montreal Neurological Institute (MNI): the template was created by averaging 305 anatomical brain scans, which had previously been linearly mapped to another template [96] (the latter had been constructed by manually aligning anatomical landmarks).

The spatial normalization was carried out using SPM8[29], where an optimum 12-parameter affine transformation was computed. However, due to the variability between individual brain anatomies, a perfect matching can never be achieved[97]. Depending on the location, inaccuracies on the order of a few centimeters are possible, which can result in diffuse group activation patterns given very localized activations on a single subject level[98].

8.4 Temporal bundling of scans

FMRI data typically exhibits a considerable degree of temporal autocorrelation[28]. The correlation is commonly regarded to be due to intrinsic physiological fluctuations. This temporal autocorrelation would create a problem if each acquired 3D volume would serve as its own observation, as both classification and permutation methods require *independent* samples (observations).

Mainly, there exist two different approaches for dealing with this problem: firstly, a specific filter can be applied to the data, which decreases the temporal autocorrelation. These filters attempt to flatten the power spectrum of the fluctuations in the fMRI data, hence the procedure is known as whitening. However the whitening procedures do not fully assure the prerequisite of full independence of the observations. An alternative strategy for dealing with the autocorrelation problem is to *concatenate* multiple observations into one observation. As the temporal range of the autocorrelations typically is on a smaller order than the time scale spanned by the concatenation, the resulting concatenated observations can be regarded as independent from each other. Most commonly, the concatenation is applied for all scans within

one experimental trial. The trial length typically is on the range of 15s to 30s, therefore given a repetition time of 2s this results in 8 to 15 scans being concatenated to one observation. There exist many variants of concatenating the scans, such as simple averaging, weighted averaging or the general linear model (GLM, see Section 4.3 on page 19). The latter method is most widely used, since it integrates the task-evoked BOLD-response properties of the brain.

More precisely, for every single experimental trial of an experiment, a custom GLM regressor in the form of a column in the design matrix X (see Equation 4.1 on page 19) is constructed. Given n voxels, m experimental conditions and with each of the conditions consisting of k trials, the design matrix X would consist of $m \cdot k$ columns. After solving for the scaling parameters β in Equation 4.3, a total of $t = m \cdot k$ observations is yielded (each observation is then compromised of a 3D volume containing m voxel-wise β -estimates, i.e. one for each trial).

Chapter 9

Multivariate Analysis & Statistics

In the following, the multivariate analysis and the nonparametric framework introduced in this thesis will be described. For the sake of overview the logic of the nonparametric framework of the single-subject analysis is displayed in Figure 9.1. The rationale of the group-level analysis is shown in Figure 9.2.

9.1 Support vector classification

For my work, I chose support vector machines (SVMs) as classifiers. SVMs have two main advantages over other classifiers, which render them as great test platforms for non-parametric analysis of fMRI data. Firstly, SVMs are extremely efficient from a computational point of view. As the nonparametric permutation-based methods used in this thesis require elaborate and hence time-intensive computations, this aspect is of essential importance. Secondly, SVMs are able to find good classification boundaries even in very high-dimensional feature spaces[99, page 93], which is the case for fMRI data. A third advantage of SVMs, which however is not used in this work but given here for the sake of completeness, is the possibility to efficiently perform a non-linear classification by mapping into a kernel space.

For my thesis I have used a *linear* SVM, albeit the classification performance may be superior for an optimized non-linear SVM. The reason for this choice is mainly its higher computational efficiency and a simpler interpretability of the results for the feature weight mapping method. Furthermore, since $n \gg t$ in fMRI data (n is the number of features, t the number of examples), the advantages of non-linear classifiers are likely rather small. Throughout this work, I use the support vector machine implementation given by the LIBSVM software package [100]. More precisely I used LIBSVM's two-class *C*-Support Vector Classification[101].

Given the data space Y (of dimension $n \times t$), the label (or indicator) space $Z = \{-1, 1\}$ and the distribution $D : Z \times Y$, a training subset is selected:

$$\{(y_1, z_1), \dots, (y_{tr}, z_{tr}) | y_i \in Y, z_i \in \{-1, 1\}\}$$

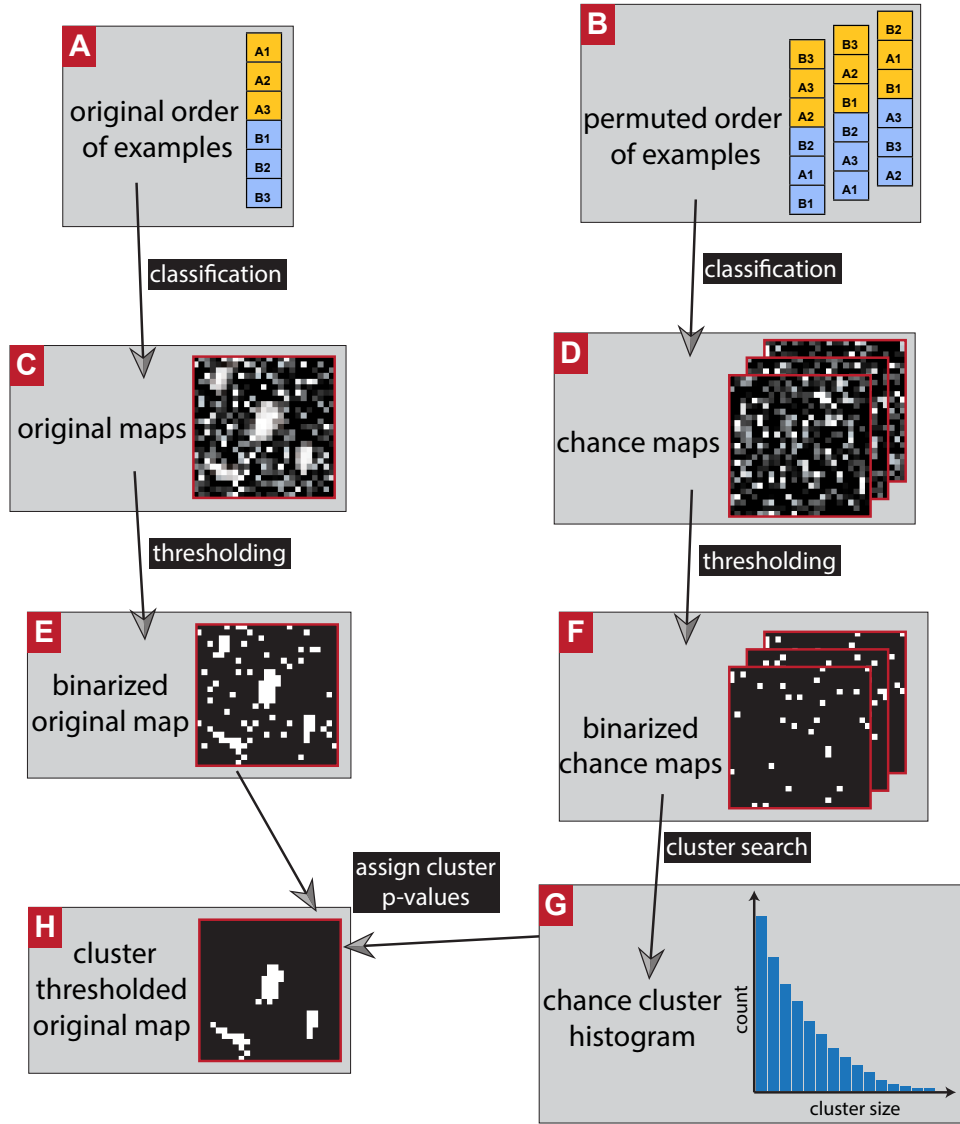


Figure 9.1: Schematic overview of the nonparametric framework on the single-subject level. **(A)** Original order of the observations. In this illustrative example, three data points are assigned to class A (in orange) and three to class B (in blue). **(B)** Permuted order of the observations. Note that for each of the permutations, the order of the observations had been shuffled randomly. Three permutations are depicted here, however, in the actual data sets, up to 10000 permutations were carried out. **(C)** Classification using the original data set yields the original information map. **(D)** Classification using the permuted data sets yields chance information maps (one for each permutation). The resulting pool of chance information maps is then used to construct a nonparametric voxel-wise threshold map (e.g. $p_{vox} = 0.001$). **(E)** Voxel-wise thresholding of the original map, only allowing voxels of the original map with a voxel-wise p -value lower than set by the voxel threshold map. All supra-threshold voxels are set to 1, while the rest of the map is set to 0. **(F)** Voxel-wise thresholding of the chance maps, only allowing voxels of the chance maps that surpass the voxel-wise threshold. The same threshold is used as before for the original map. All supra-threshold voxels are set to 1, while the rest of the map is set to 0. **(G)** Cluster search in the binarized chance maps, yielding a cluster-size record. **(H)** A cluster search within the original binarized map yields a list of cluster sizes. Cluster-level p -values can be assigned to this list by usage of the chance cluster size record of step G, allowing an application of a cluster-level threshold (including a FDR correction on the cluster p -values).

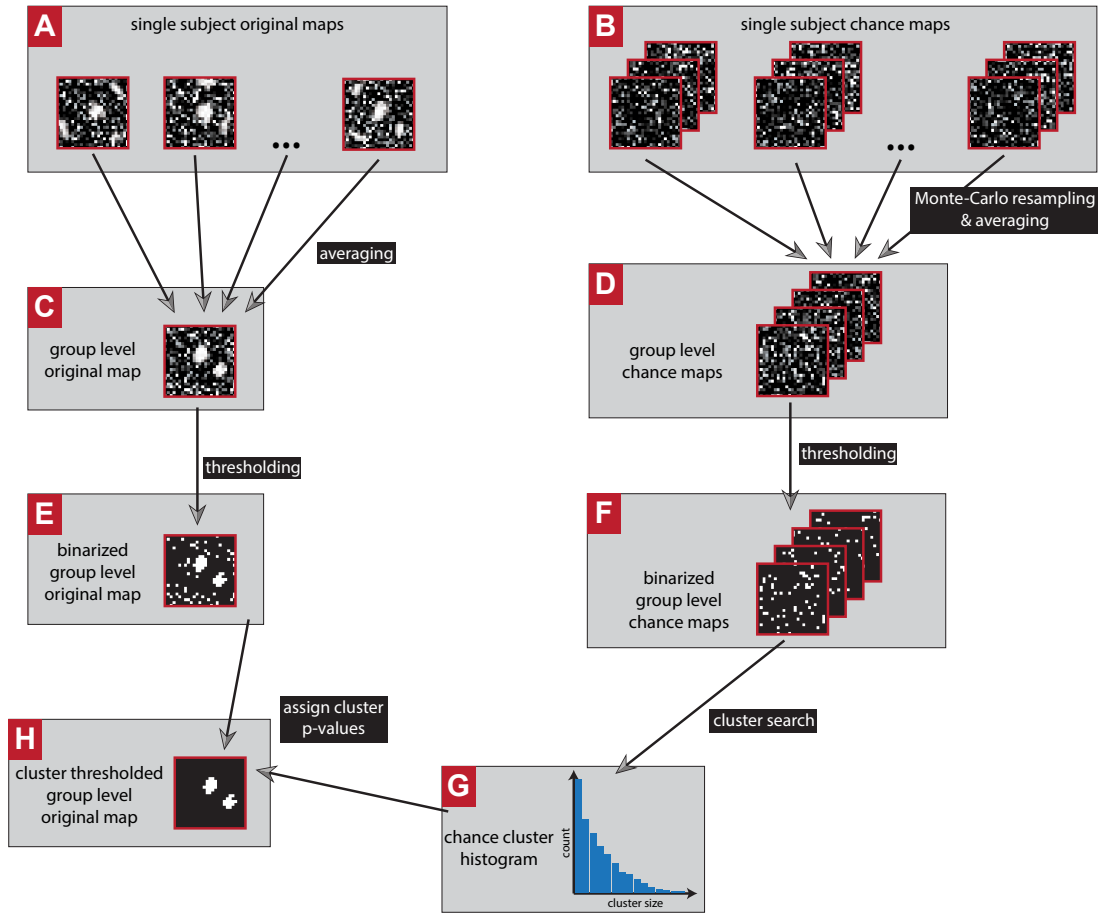


Figure 9.2: Schematic overview of the nonparametric framework on the group-level. **(A)** Classification using the original data set yields the original information maps, one for each subject. **(B)** Classification using the permuted data sets yields chance information maps for of the subjects (one for each permutation). For the group-level analysis, 100 permutations per subject were used. **(C)** Averaging of the single-subject original maps, yielding one group map. **(D)** Monte-Carlo resampling procedure, one permutation of each subject is selected at random and the selected maps are averaged. The procedure is carried out for 10^5 repetitions, yielding 10^5 group chance maps. The resulting pool of chance group information maps is then used to construct a nonparametric voxel-wise threshold map (e.g. $p_{vox} = 0.001$). **(E)** Voxel-wise thresholding of the original group map, only allowing voxels of the original map with a voxel-wise p -value lower than set by the voxel threshold. All supra-threshold voxels are set to 1, while the rest of the map is set to 0. **(F)** Voxel-wise thresholding of the group chance maps, only allowing voxels of the chance maps that surpass the voxel-wise threshold. The same threshold is used as before for the original map. All supra-threshold voxels are set to 1, while the rest of the map is set to 0. **(G)** Cluster search in the binarized group chance maps, yielding a cluster-size record. **(H)** A cluster search within the original binarized group map yields a list of cluster sizes. Cluster-level p -values can be assigned to this list by usage of the chance cluster size record of step *G*, allowing an application of a cluster-level threshold (including a FDR correction on the cluster p -values).

where $t_{tr} < t$. In the case of a leave-one-out cross-validation (LOOCV) scheme, for each cross-validation step two examples are used as test set and the rest as training data, hence $t_{tr} = t - 2$. The C -SVM solves the following problem, in finding the minimum for [101]

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^{t_{tr}} \xi_i \quad (9.1)$$

so that the constraint

$$z_i(\mathbf{w}^T \mathbf{y}_i + b) \geq 1 - \xi_i \quad (9.2)$$

is fulfilled for all slack variables $\xi_i \geq 0$ and for all $i = 1, \dots, t_{tr}$. The positive parameter C is a regularization parameter here, which penalized non-zero slack variables ξ_i . Effectively, these slack variables allow misclassification of training examples, therefore the C -SVM is referred to as a *soft margin* classifier.

It should be noted that for computational reasons, SVM algorithms do not solve the primal problem of equation 9.1 but rather in the dual form, which gives equivalent solutions to the problem in its primal form. Furthermore, it should be mentioned that prior to classification, each column of the data matrix Y had been linearly scaled (by means of multiplication and addition of an offset), so that the data falls in the interval $[-1, +1]$.

The result of the (linear) classifier training is a training weight vector \vec{w} . Given the data of an experimental trial, which is a vector of features y , with $y \in Y$ (i.e. a row in the data matrix Y), the class membership is derived by the following: if the dot product between the weight and feature vector $\vec{w} \cdot \vec{y} > 0$, then the trial is classified as belonging to class A , while class B is determined if the product is negative. The training vector \vec{w} can then be used for unseen data points of the test set. As the labels of the test set are known, the predicted labels of the classifier can be compared to the known labels.

9.2 Searchlight decoding (SLD)

For searchlight decoding, the standard volumetric spherical searchlight approach [63, 102] was used. The diameter d of the searchlight was set to five voxels. This corresponded to a physical diameter of 1.5mm in case of a voxel size of 3mm. For each location within the brain, all voxels within the searchlight centered at the location were extracted (for all of the t observations). In case of the fMRI experiments, these observations were the three-dimensional β -estimate maps. This spatial selection of voxels was put into a linear support vector machine (see Section 9.1 on page 65), performing a $t/2$ -fold LOOCV procedure (see Section 9.1 and Section 4.4.4 on page 26). In short, each cross-validation step used $t - 2$ observations ($t/2 - 1$ from each condition) as a training set and two observations (one from each condition) as a test set. Over the course of $t/2$ cross-validation folds, the classifier was trained on the training set and the labels of the

unseen test set were predicted. The average accuracy (i.e., the percentage of correctly predicted labels) of the $t/2$ cross-validation steps was then mapped onto the center of the location. The procedure was applied to all locations of a whole brain mask and resulted in a map A of decoding accuracies [5].

9.3 Feature weight mapping (FWM)

In the feature weight mapping method, each feature dimension is assigned a *weight*, which is directly derived from the underlying classifier's mathematical model (in this thesis a linear SVM). The weight of a feature is an indicator of the *contribution* to the classification decision and can thus be interpreted as the feature importance. More precisely, training a linear classifier on a the full data set yields a weight vector \vec{w} , which was derived by the classifier's training (see Section 9.1).

As the number of features of the data matrix Y is very high in the case of fMRI, the classification performance is likely inferior. For this reason, a dimensionality reduction procedure is applied prior to classification. For my thesis, the principal component analysis (PCA) method is used for this reduction. The idea behind PCA is to compress the information from Y while maximizing the information content. The PCA procedure obtains a new representation Y^* of the matrix as a linear combination of the columns (features) of Y , so that the first principal component (i.e. the first row of Y^*) contains the largest part of information about the original data matrix Y [103]. The maximal number of principal components is equal to the number of observations, i.e. the number of rows in the original data matrix Y . For the computation of the PCA projection, the singular value decomposition of Y is obtained [103]:

$$Y = P\Delta Q^T \quad (9.3)$$

where P is the matrix of left singular values, Q the matrix of right singular values and Δ is the diagonal matrix of singular values, which is equivalent to the nonzero eigenvalues of $Y^T Y$. As the values of Y are real numbers, the matrices P and Q can be regarded as rotation matrices.

The transformation of the data matrix Y into the PCA space Y^* is then given by [103]:

$$Y^* = YQ \quad (9.4)$$

and similarly, the transformation from the PCA space Y^* into the voxel space is given by

$$Y = Y^*Q^T \quad (9.5)$$

9.4 Permutation testing

On the single-subject level, permutation tests were employed. The starting point for this was the data space Y of the dimension $t \times n$, where t is the number of examples and n the number of features. Note that Y is not the entire data space of an fMRI experiment or a simulation, but either a spatial preselection of voxels (SLD method), or the dimensionality-reduced data space (FWM method). The permutation is carried out by randomly shuffling the *order* of the examples, i.e. by interchanging the rows of Y .

It should be highlighted that the permutation of the rows was carried out before the data set was split into training or test sets. This ensured that there is no bias due to an uneven class distribution in the test or training sets [5]. In case of the SLD method, one permutation was used for a full searchlight decoding on all locations. This ensured the preservation of the spatial correlations present in the data. As a result of one single permutation, one *chance* decoding accuracy map \tilde{A} was created. For the FWM method, the permutations were computed in the PCA space, i.e. the rows of the matrix Y^* were shuffled. After the computation of the permutations, the resulting chance weights were projected back into the voxel space. Hence the result of one single permutation procedure was a chance weight map \tilde{W} in voxel space.

Depending on the type of analysis in terms of single-subject or group studies, a different number of permutations was used: in case of single-subject analysis, up to 10^4 permutations were computed, while 10^2 permutations per subject were sufficient for a group-level analysis.

9.5 Group level Monte-Carlo recombination

For the group analysis, an intermediate step had to be carried out which combined the results of single-subject analysis on the group maps (this step was hence unnecessary for a single-subject analysis). In case of SLD these were the chance accuracy maps $\tilde{A}_{i,j}$ (i denotes the index of the accuracy map, i.e. a number between 1 and 10^2 , j denotes the subject), in case of the FWM method the single-subject maps were chance weight maps $\tilde{W}_{i,j}$ (i stands for index of the weight map, i.e. a number between 1 and 10^2 , j denotes the subject). The group recombination procedure was identical for the SLD and FWM methods and was based on 100 permutations for each of the N_{sub} subjects. Then, for each subject, one out of the 100 chance maps from the permutation procedure was drawn randomly (with replacement). Next, this selection of N_{sub} chance maps was averaged voxel-wise, yielding *one* chance map *on group-level*, i.e. one chance group accuracy map \tilde{F} (SLD) or one chance group weight map \tilde{G} (FWM). The procedure of random selection and averaging was repeated 10^5 times, which resulted in a large pool of chance group maps \tilde{F}_m or \tilde{G}_m , $m = 1 \dots 10^5$.

The averaging procedure was also carried out on the original (not permuted) accuracy or weight maps. For this, in case of the SLD method, the accuracy maps A_j were averaged over all N_{sub} subjects, resulting in a group accuracy map F . Equivalently, in case of the FWM method, the weight maps W_j were averaged across the N_{sub} subjects, which resulted in a group weight map G .

9.6 Threshold map procedure

As a prerequisite for the subsequent cluster-based analysis, an empirical voxel-wise *chance* distribution is required. In the case of single-subject analysis, the permuted chance maps (\tilde{A}_i in SLD or \tilde{W}_i in FWM; $i = 1 \dots N_{perms}$) were used directly for constructing this voxel-wise distribution. In the case of group studies, the recombined group chance maps were utilized (\tilde{F}_i in SLD or \tilde{G}_i in FWM; $i = 1 \dots 10^5$). For the threshold procedure, the voxel-wise histogram of chance values was computed first. This allowed the determination of the threshold value, which was either an accuracy value (SLD) or a feature weight (FWM). In the case of accuracy maps, the accuracy for which the right-tailed area of the normalized histogram of voxel-wise accuracies was below p_{vox} was determined. In the case of training weights, both a left-tailed and a right-tailed threshold weight was determined, since the voxel-wise training weight could take both negative and positive values. In the latter case, the voxel-threshold p_{vox} was divided by a factor of two, so that the right-tailed and left-tailed area taken together was equally large as the the right-tail area in the accuracy maps of the SLD method. The procedure was repeated for all voxels, yielding one threshold map T (SLD) or two threshold maps T^+ and T^- (FWM). Hence, in a statistical notion, the threshold maps represent the accuracy or weight level which, if surpassed, would label the voxel as being less probable than p_{vox} .

Next, the threshold maps were used to binarize both the original maps and chance maps. If a voxel surpassed the threshold, it was set to 1, otherwise it was set to 0. For the FWM method and the negative weight map, the voxel in the binary image was set to 1 if it was below the threshold.

In case of the SLD method, this resulted in a binarized original accuracy map B and a pool of binarized chance accuracy maps \tilde{B}_i where $i = 1 \dots N_{perms}$ for a single-subject study and $i = 1 \dots 10^5$ for a group-level study. In case of the FWM method, two binarized weight maps for positive and negative weights C^+ and C^- were the result from the thresholding procedure. In analogy to the SLD method, a pool of binarized chance weight maps \tilde{C}_i^+ and \tilde{C}_i^- were created, where $i = 1 \dots N_{perms}$ for a single-subject study and $i = 1 \dots 10^5$ for a group-level study.

9.7 Cluster size statistics

Given the binarized chance maps (\tilde{B}_i in case of SLD and \tilde{C}_i^+ respectively \tilde{C}_i^- in case of FWM), it was possible to investigate the spatial features of these maps in terms of connected components, i.e. *clusters*. To perform a cluster search, I used a 6-connectivity scheme; two nonzero voxels were considered connected if they shared a face, but not if they only shared an edge or vertex. In other words, in the example of SLD, a voxel was joined to a cluster only if its accuracy exceeded the accuracy corresponding to the p_{vox} in the threshold map.

I applied the cluster search using the above algorithm either for the N_{perms} permutation maps directly in the case of single-subject studies to the or in the 10^5 recombined group maps. The occurring cluster sizes were collected in form of a cluster list (\tilde{L}_{cl} in case of SLD or two cluster lists, one for negative and one for positive weights; \tilde{L}_{cl}^+ and \tilde{L}_{cl}^- in case of FWM).

For each of the chance maps *all* occurring cluster sizes were recorded (the minimum size for a cluster was two voxels). This allowed the computation of an empirical cluster size distribution H_{cl} (SLD) or H_{cl}^+ and H_{cl}^- (FWM), which was defined as the normalized histogram of cluster sizes in the chance record:

$$H_{cl} = \frac{N_s}{\sum_{s=1}^{max(L_{cl})} N_s} \quad (9.6)$$

where N_s is the occurrence of a cluster of the size s and $max(L_{cl})$ the largest cluster detected.

Critically, the same cluster search was applied to the original, binarized accuracy map B or weight maps C^+ and C^- and the present cluster sizes were gathered into a list L_{cl} (SLD) or L_{cl}^+ and L_{cl}^- in case of FWM (for clusters with positive and negative weight). Given this cluster size record L_{cl} and the empirical chance cluster size distribution H_{cl} , it is possible to compute the probability for the occurrence of a discovered cluster size in the original data: a cluster with the size s is computed to have a p -value of

$$p_{cl} = \sum_{s' \geq s}^{max(L_{cl})} H_{cl}(s') \quad (9.7)$$

Hence, it is possible to assign a p -value to each cluster size and introduce a threshold of cluster size for reaching significance. For the FWM method, p -values were computed separately for positive and negative weights.

To correct for multiple comparisons at cluster level, I implemented a step-down FDR method (as introduced in Section 5.4.2 on page 43) for the list P_{cl} of all cluster p -values of the original map. All clusters with a probability $p_{cl} > 0.05$ were discarded, which yielded cluster-size controlled accuracy or weight maps.

9.8 Parametric framework for comparison

The proposed nonparametric statistical framework was compared with T-tests, which are the most commonly practiced parametrical alternative[102, 104, 105]. All T-based analysis was carried out in SPM8[29]. In case of searchlight decoding (SLD), a one-tailed T-test against the theoretical chance level of 0.5 was carried out. In the case of classification weight mapping (FWM), a two-tailed T-test against 0 was carried out. The resulting T-maps were thresholded with a voxel-wise p -value of p_{vox} . For the two-tailed T-test, two one-tailed T-tests were carried out, each thresholded at $\frac{p_{vox}}{2}$ (which effectively corresponds to a two-tailed tests at p_{vox}). This procedure ensured full comparability to procedure of two threshold maps in the feature weight mapping method for the nonparametric framework. A multiple comparisons correction using Gaussian random field methods (see Section 5.4.3 on page 44) allowed the derivation of p -values for the resulting supra-threshold clusters (using SPM8[29]). These p -values were then

corrected using either standard false-discovery rate based (FDR, see Section 5.4.2 on page 43) or Bonferroni based methods (Familywise error FWE correction, see Section 5.4.1 on page 43) as implemented in SPM8.

9.9 Processing pipelines

In the following, I will summarize the processing pipelines for the SLD and FWM method, both on the single-subject level and group-level. As a zeroth step of each pipeline, preprocessing¹ (motion correction, temporal filtering, spatial normalization to standard brain space, temporal bundling resulting in one β -estimate map per trial) was carried out. This resulted in a data matrix Y consisting of N_{vox} voxels and t examples for each of the N_{sub} subjects.

9.9.1 SLD on single subject level

1. Searchlight procedure, extraction of a spherical neighborhood of diameter d for every location k , yielding a searchlight data matrix $Y'(k)$ for every voxel $k = 1 \dots N_{vox}$. For each searchlight location, a leave-one-out cross-validation support vector classification was computed. The mean percentage of correctly identified labels of the test set was copied into the accuracy map matrix $A(k)$, i.e. to the location corresponding to the searchlight's center voxel.
2. Permuted searchlight procedure, extraction of a spherical neighborhood of diameter d for every location k , yielding a searchlight data matrix $Y'(k)$ for every voxel $k = 1 \dots N_{vox}$ and $j = 1 \dots N_{sub}$. For each searchlight location, a leave-one-out cross-validation support vector classification was performed 10^4 times with permuted order of the rows of $Y'(k)$. For each permutation $i = 1 \dots 10^4$, the mean percentage of correctly identified labels of the test set was copied into the chance accuracy map matrices to the location corresponding to the searchlight's center voxel $\tilde{A}_i(k)$, $i = 1 \dots 10^4$. Importantly, each permuted order i was held fixed for all searchlight locations.
3. Threshold map procedure, computation of a voxel-wise histogram of the permuted accuracy maps \tilde{A}_i , $i = 1 \dots 10^4$, determining the threshold map T according to the right-sided accuracy value which corresponds to a probability of p_{vox} or smaller.
4. Binarization of the original accuracy map A . If the k -th voxel $A(k)$ surpasses the threshold set by $T(k)$ then $B(k) = 1$, otherwise $B(k) = 0$. The procedure is repeated for all voxels, $k = 1 \dots N_{vox}$.
5. Binarization of the permuted accuracy maps \tilde{A}_i . If the k -th voxel $\tilde{A}_i(k)$ surpasses the threshold set by $T(k)$ then $\tilde{B}_i(k) = 1$, otherwise $\tilde{B}_i(k) = 0$. The procedure is repeated for all voxels, $k = 1 \dots N_{vox}$ and all permutations $i = 1 \dots 10^4$.
6. Cluster search in the binarized accuracy map B , resulting in cluster list L_{cl}
7. Cluster search in the binarized chance maps \tilde{B}_i with $i = 1 \dots 10^4$. This resulted in a cluster list \tilde{L}_{cl}

¹this step was omitted for all simulations. For the 7T ultra-high resolution data set no spatial normalization to MNI space was carried out.

8. Computation of the histogram of \tilde{L}_{cl} , resulting in H_{cl}
9. Derivation of cluster-wise p -values for all clusters in the (not permuted) cluster size list L_{cl} using chance cluster size histogram H_{cl} . This resulted in a list of p -values P_{cl}
10. Step-down FDR-correction of the p -value list P_{cl} and thresholding of original accuracy map A , only allowing clusters with a corrected cluster p -value smaller than 0.05.

9.9.2 SLD on the group level

1. Searchlight procedure, extraction of a spherical neighborhood of diameter d for every location k , yielding a searchlight data matrix $Y_j'(k)$ for every voxel $k = 1 \dots N_{vox}$ and all subjects $j = 1 \dots N_{sub}$. For each searchlight location, a leave-one-out cross-validation support vector classification was computed and the mean percentage of correctly identified labels of the test set was copied into the accuracy map matrix $A_j(k)$, i.e. to the location corresponding to the searchlight's center voxel. The procedure was repeated for all subjects $j = 1 \dots N_{sub}$.
2. Permuted searchlight procedure, extraction of a spherical neighborhood of diameter d for every location k , yielding a searchlight data matrix with shuffled rows $\tilde{Y}_{i,j}'(k)$ for every voxel $k = 1 \dots N_{vox}$ and every subject $j = 1 \dots N_{sub}$ (with the permutation index $i = 1 \dots 10^2$). For each searchlight and permutation i , a leave-one-out cross-validation support vector classification was performed. The mean percentage of correctly identified labels of the test set was copied into the chance accuracy map matrices to the location corresponding to the searchlight's center voxel $\tilde{A}_{i,j}(k)$, $i = 1 \dots 10^2$. Importantly, the permutation was held fixed for all locations $k = 1 \dots N_{vox}$. The procedure was repeated for all subjects $j = 1 \dots N_{sub}$.
3. Averaging of A_j with $j = 1 \dots N_{sub}$ over all subjects, resulting in mean group accuracy map F
4. Monte-Carlo group resampling procedure: random selection of one permuted accuracy map per subject $\tilde{A}_{i,j}$, $i \in [1, 100]$, $j = 1 \dots N_{sub}$ and averaging of these N_{sub} maps. The step was repeated for 10^5 times, resulting in chance group accuracy maps \tilde{F}_m with $m = 1 \dots 10^5$
5. Threshold map procedure, computation of a voxel-wise histogram of the permuted group accuracy maps \tilde{F}_m , $m = 1 \dots 10^5$. This determines the threshold map T according to the right-sided accuracy value which corresponds to a probability of p_{vox} or smaller.
6. Binarization of the original group accuracy map F . If the k -th voxel $F(k)$ surpasses the threshold set by $T(k)$ then $B(k) = 1$, otherwise $B(k) = 0$. The procedure is repeated for all voxels, $k = 1 \dots N_{vox}$.
7. Binarization of the chance group accuracy maps \tilde{F}_m . If the k -th voxel $\tilde{F}_i(k)$ surpasses the threshold set by $T(k)$ then $\tilde{B}_i(k) = 1$, otherwise $\tilde{B}_i(k) = 0$. The procedure is repeated for all voxels, $k = 1 \dots N_{vox}$ and all resampling steps $m = 1 \dots 10^5$.
8. Cluster search in the binarized group accuracy map B , resulting in cluster list L_{cl}
9. Cluster search in the binarized group chance maps \tilde{B}_m with $m = 1 \dots 10^5$. This resulted in the cluster list \tilde{L}_{cl}
10. Computation of the histogram of \tilde{L}_{cl} , resulting in H_{cl}

11. Derivation of cluster-wise p -values for all clusters in the (not permuted) cluster size list L_{cl} using chance cluster size histogram H_{cl} . This resulted in a list of p -values P_{cl}
12. Step-down FDR-correction of the p -value list P_{cl} and thresholding of original group accuracy map F , only allowing clusters with a corrected cluster p -value smaller than 0.05.

9.9.3 FWM on single subject level

1. Projection of the data matrix Y along its principal components, resulting in the matrix of reduced dimension Y^*
2. Support vector classification of Y^* , resulting in a weight map W^*
3. Support vector classification of Y^* using a random permutations of the rows of Y^* . Step repeated for 10^4 times, resulting in \widetilde{W}_i^* with $i = 1 \dots 10^4$.
4. Back-projection of W^* into the voxel space, resulting in weight map W .
5. Back-projection of \widetilde{W}_i^* into the voxel space, resulting in permuted weight maps \widetilde{W}_i with $i = 1 \dots 10^4$.
6. Threshold map procedure, computation of a voxel-wise histogram of the permuted weight maps \widetilde{W}_i , $i = 1 \dots 10^4$, determining the threshold maps T^+ and T^- according to the right-sided and left-sided weight value which corresponds to a probability of $\frac{p_{vox}}{2}$.
7. Binarization of the original weight map W . If the k -th voxel $W(k)$ *surpasses* the threshold set by $T^+(k)$ then $C^+(k) = 1$, otherwise $C^+(k) = 0$. If the voxel $W(k)$ *falls below* the threshold set by $T^-(k)$ then $C^-(k) = 1$, otherwise $C^-(k) = 0$. The procedure is repeated for all voxels, $k = 1 \dots N$.
8. Binarization of the permuted weight maps \widetilde{W}_i . If the k -th voxel $\widetilde{W}_i(k)$ *surpasses* the threshold set by $T^+(k)$ then $\widetilde{C}_i^+(k) = 1$, otherwise $\widetilde{C}_i^+(k) = 0$. If the voxel $\widetilde{W}_i(k)$ *falls below* the threshold set by $T^-(k)$ then $\widetilde{C}_i^-(k) = 1$, otherwise $\widetilde{C}_i^-(k) = 0$. The procedure is repeated for all voxels, $k = 1 \dots N$ and all permutations $i = 1 \dots 10^4$.
9. Cluster search on the binarized weight map C^+ , resulting in the cluster list L_{cl}^+ and separately a cluster search on C^- , resulting in the cluster list L_{cl}^-
10. Cluster search in the binarized chance maps \widetilde{C}_i^+ , $i = 1 \dots 10^4$, resulting in the cluster list \widetilde{L}_{cl}^+ and separately a cluster search on \widetilde{C}_i^- , resulting in the cluster list \widetilde{L}_{cl}^-
11. Computation of the histogram of \widetilde{L}_{cl}^+ , resulting in H_{cl}^+ . Separate computation of the histogram of \widetilde{L}_{cl}^- , resulting in H_{cl}^- .
12. Derivation of cluster-wise p -values for all clusters in the (not permuted) cluster size list L_{cl}^+ using chance cluster size histogram H_{cl}^+ , resulting in a list of p -values P_{cl}^+ . Separate derivation of p -values for all clusters in L_{cl}^- using the chance cluster size histogram H_{cl}^- , resulting in a list of p -values P_{cl}^-

13. Step-down FDR-correction of both p -value lists P_{cl}^+ and P_{cl}^- and thresholding of original weight map W , only allowing clusters with a corrected cluster p -value smaller than 0.05.

9.9.4 FWM on the group level

1. Projection of the data matrix Y_j corresponding to subject j along it's principal components, resulting in a matrix of reduced dimension Y_j^* , procedure repeated for all subjects $j = 1 \dots N_{sub}$.
2. Support vector classification of Y_j^* , resulting in a weight map W_j^* , procedure repeated for all subjects $j = 1 \dots N_{sub}$.
3. Support vector classification of $\tilde{Y}_{i,j}^*$ using a random permutations of the rows. Step repeated for $i = 1 \dots 10^2$, resulting in $\tilde{W}_{i,j}^*$, $i = 1 \dots 10^2$, for all subjects $j = 1 \dots N_{sub}$.
4. Back-projection of W_j^* into the voxel space, resulting in weight map W_j , $j = 1 \dots N_{sub}$.
5. Back-projection of $\tilde{W}_{i,j}^*$ into the voxel space, resulting in permuted weight maps $\tilde{W}_{i,j}$, $i = 1 \dots 10^2$, $j = 1 \dots N_{sub}$.
6. Averaging over all subjects of the original weight maps W_j , $j = 1 \dots N_{sub}$. This results in a mean group weight map G .
7. Monte-Carlo group resampling procedure: random selection of one permuted weight map per subject $\tilde{W}_{i,j}$, $i \in [1, 100]$, $j = 1 \dots N_{sub}$ and averaging of the N_{sub} maps. The step was repeated for 10^5 times, resulting in chance group weight maps \tilde{G}_m , $m = 1 \dots 10^5$.
8. Threshold map procedure, computation of a voxel-wise histogram of the permuted weight maps \tilde{G}_m , $m = 1 \dots 10^5$. This determines the threshold maps T^+ and T^- according to the right-sided and left-sided weight value which correspond to a probability of $\frac{p_{vox}}{2}$.
9. Binarization of the original group weight map G . If the k -th voxel $G(k)$ *surpasses* the threshold set by $T^+(k)$, then $C^+(k) = 1$, otherwise $C^+(k) = 0$. If the voxel $G(k)$ *falls below* the threshold set by $T^-(k)$, then $C^-(k) = 1$, otherwise $C^-(k) = 0$. The procedure is repeated for all voxels, $k = 1 \dots N_{vox}$.
10. Binarization of the permuted weight maps \tilde{G}_m . If the k -th voxel $\tilde{G}_m(k)$ *surpasses* the threshold set by $T^+(k)$, then $\tilde{C}_i^+(k) = 1$, otherwise $\tilde{C}_i^+(k) = 0$. If the voxel $\tilde{G}_m(k)$ *falls below* the threshold set by $T^-(k)$, then $\tilde{C}_i^-(k) = 1$, otherwise $\tilde{C}_i^-(k) = 0$. The procedure is repeated for all voxels, $k = 1 \dots N_{vox}$ and all resampling steps $m = 1 \dots 10^5$.
11. Cluster search in the binarized weight map C^+ , resulting in the cluster list L_{cl}^+ and separately cluster search on C^- , resulting in the cluster list L_{cl}^- .
12. Cluster search on the binarized chance maps \tilde{C}_m^+ , resulting in the cluster list \tilde{L}_{cl}^+ and separately a cluster search on \tilde{C}_m^- , resulting in cluster list \tilde{L}_{cl}^- for all resampling steps $m = 1 \dots 10^5$.
13. Computation of the histogram of \tilde{L}_{cl}^+ , resulting in H_{cl}^+ . Separate computation of the histogram of \tilde{L}_{cl}^- , resulting in H_{cl}^- .

14. Derivation of cluster-wise p -values for all clusters in the (not permuted) cluster size list L_{cl}^+ using the chance cluster size histogram H_{cl}^+ resulting in a list of p -values P_{cl}^+ . Separate derivation of p -values for all clusters in L_{cl}^- using chance cluster-size histogram H_{cl}^- , resulting in a list of p -values P_{cl}^-
15. Step-down FDR-correction of both p -value lists P_{cl}^+ and P_{cl}^- and thresholding of original weight map W , only allowing clusters with a corrected cluster p -value smaller than 0.05.

Part III

Results

Chapter 10

Singe subject results

10.1 Single subject geometric simulation

The goal of this simulation was to investigate the behavior of the feature weight mapping (FWM) method in comparison with the searchlight decoding (SLD) method, both analyzed by the proposed nonparametric framework. The simulation exhibits a precise deposition of information at pre-specified areas (see Figure 10.1A). Ultimately, this allowed the ability to quantitatively compare the SLD and FWM methods in dependency of the underlying geometry of the distribution of class information. In total, three levels of the graininess of information distribution were used here (the leftmost, middle and rightmost half-cubes in Figure 10.1A, representing fine, medium and coarse information distribution).

10.1.1 Qualitative comparison between the FWM and SLD method

Figure 10.1 allows a qualitative comparison between the feature weight mapping and searchlight decoding methods. The information distribution is depicted in Figure 10.1A, the violet areas stand for informative regions of condition A , while the blue areas stand for informative regions of condition B . Both methods implemented the proposed nonparametric statistics, based on random permutations and cluster size control. As the results highly depend on the voxel-wise threshold, each method was computed using two voxel-wise thresholds.

For the low threshold ($p_{vox} = 0.05$ one-sided in case of SLD and two-sided for the FWM method), the SLD method labels most informative regions as significant, while also labeling a considerable number of voxels outside the informative regions significant (Figure 10.1B). The effect becomes especially predominant in the leftmost and middle thirds of the figure, representing fine and medium information distribution. Here, the SLD method appears to overestimate the local information content. In contrast, the FWM method delineates the informative regions with a high precision (see Figure 10.1C), and does not label voxels outside of the informative regions as significant. The number of true positives, however, was smaller as compared to the SLD method, as not all informative voxels are declared significant here.

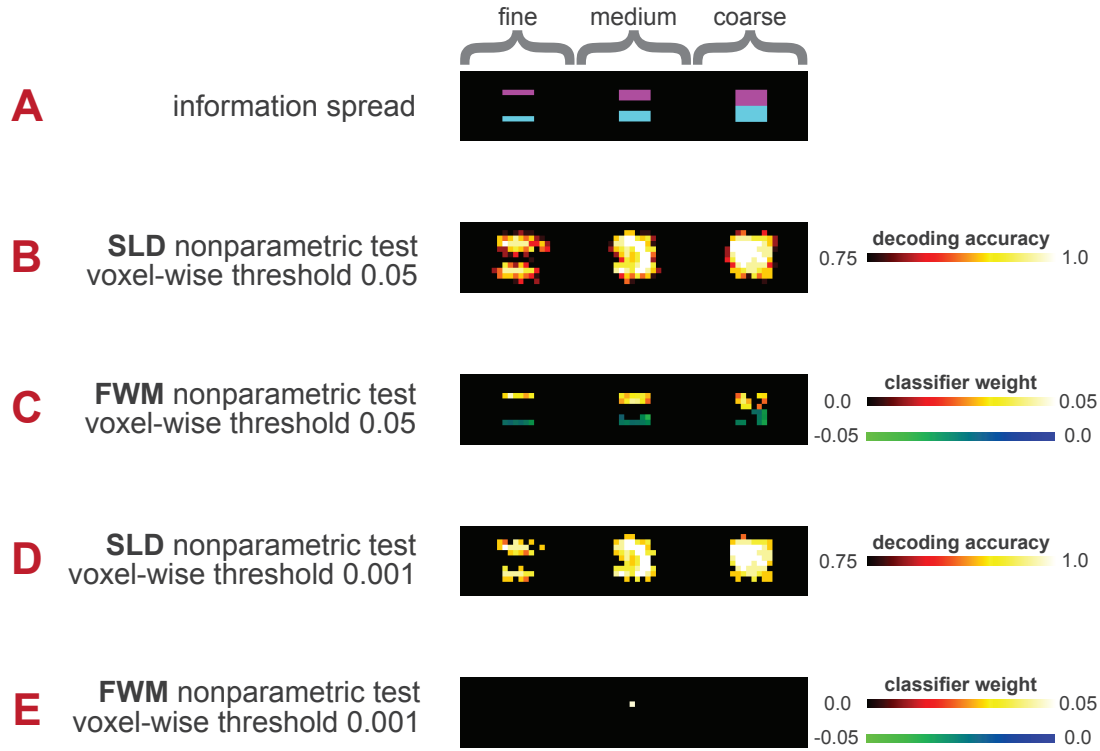


Figure 10.1: Overview of the single-subject geometric simulation results. (A) Distribution of information, the three violet half-cubes contained class information for condition *A*, the three blue half-cubes contained class information for class *B*. In total, three distinct levels of geometry of information distribution were available, the leftmost half-cubes represented a fine information spread, the middle ones an intermediate level and the rightmost half-cubes a coarse information spread. (B) Results of the searchlight decoding method using the low threshold. The results were corrected with the proposed nonparametric framework, using a voxel-wise threshold of $p_{vox} = 0.05$. (C) Results of the feature weight mapping method using the low threshold. The blue-green colors stand for negative weights, the red colors for positive weights. The results implement the nonparametric multiple comparisons correction proposed in this thesis. The voxel-wise threshold was set to $p_{vox} = 0.05$ (two-sided) (D) Results for the SLD method, using a voxel-wise threshold of $p_{vox} = 0.001$ (E) Results for the FWM method, using a two-sided voxel-wise threshold of $p_{vox} = 0.001$

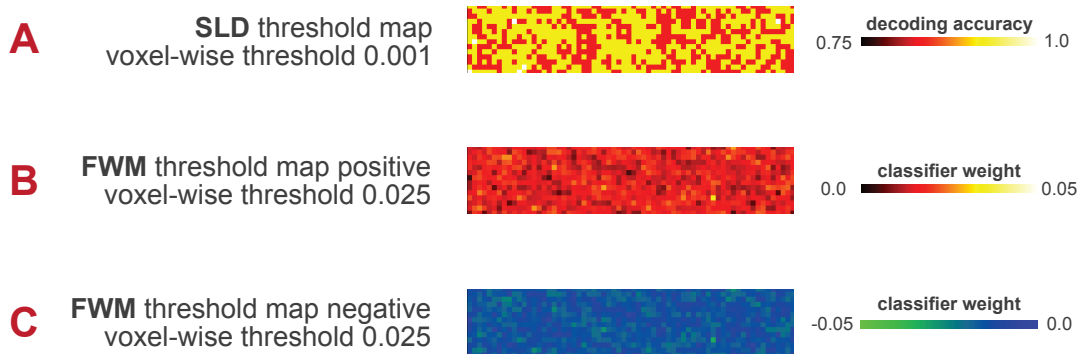


Figure 10.2: Threshold maps for the single-subject geometric simulation using the optimal voxel-wise thresholds for both SLD and FWM. The optimal thresholds are determined on a heuristic basis (in the following section). (A) The optimal voxel-wise threshold for SLD corresponds to $p_{vox} = 0.001$. As the decoding accuracies are of discrete nature since, the map displays only two values here. (B) For the FWM method and one-sided $p_{vox} = 0.025$ for positive weights was used (which corresponds to a two-sided voxel-wise threshold of $p_{vox} = 0.05$). (C) For negative weight in the CWM method a one-sided $p_{vox} = 0.025$ was used (corresponding to a two-sided voxel-wise threshold of $p_{vox} = 0.05$).

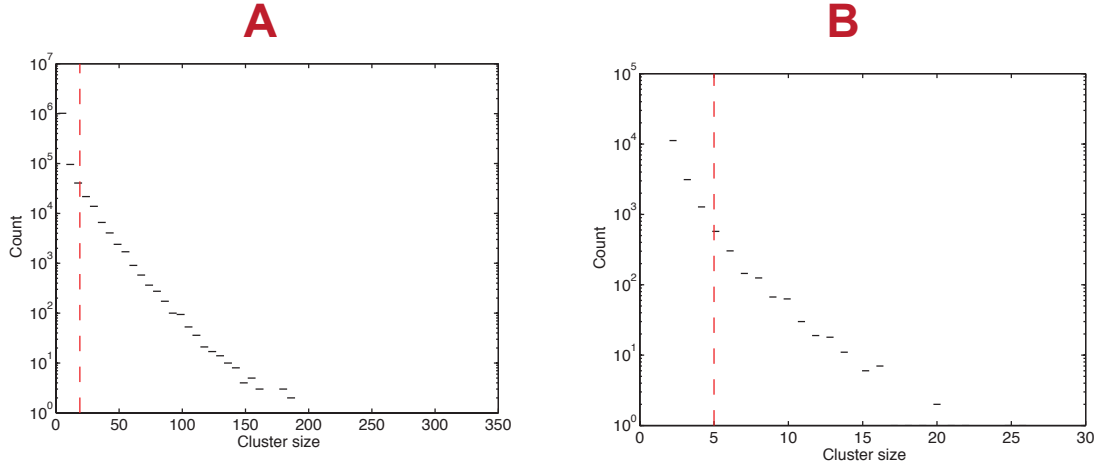


Figure 10.3: Cluster-size histogram for the single-subject geometric simulation analyzed by the SLD method. The red line in the histograms marks the cluster size corresponding to the (uncorrected) cluster p -value $p_{cl} = 0.05$. **(A)** Cluster-size histogram using a voxel-wise threshold of $p_{vox} = 0.05$. Clusters with a size larger than 19 voxels obtain a p -value $p_{cl} < 0.05$ here **(B)** Cluster-size histogram implementing a voxel-wise threshold with $p_{vox} = 0.001$. The cluster size corresponding to the cluster p -value $p_{cl} = 0.05$ was 5 voxels

Using the high threshold ($p_{vox} = 0.001$), the SLD results appear improved, as the number of significant voxels outside the informative regions was smaller than for the low threshold. On the other hand, using the high threshold the FWM method leads to almost no voxels labeled as significant. A more in-depth quantitative analysis of the impact of the voxel-wise threshold in dependence of the geometry of information distribution is found in [Section 10.1.2 on page 86](#).

The threshold maps for the nonparametric framework are displayed in [Figure 10.2](#) for both the SLD and FWM method. Note that the voxel-wise thresholds displayed here differ for both methods, the threshold was set to $p_{vox} = 0.001$ for SLD and a two-sided threshold of $p_{vox} = 0.05$ for the FWM method (which corresponds to two one-sided thresholds $p_{vox} = 0.025$ for positive and negative weights respectively). Furthermore it should be stated that in the case of the SLD method, there exists only 31 possible values for the decoding accuracy (as each condition consisted of 15 examples), therefore the threshold map was very coarse.

The empirical cluster size histograms are depicted in [Figure 10.3](#) (SLD method) and [Figure 10.4](#) (FWM method). For the SLD method and the low voxel-wise threshold of $p_{vox} = 0.05$, the minimum cluster size for obtaining a cluster p -value of $p_{cl} < 0.05$ was 19 voxels; for the higher voxel threshold $p_{vox} = 0.001$ the minimum size for the SLD method was 3 voxels. In case of the FWM method and the low threshold ($p_{vox} = 0.05$ two-sided), the minimum size for $p_{cl} = 0.05$ was 7 voxels (both for positive and negative weights). For the high voxel-wise threshold $p_{vox} = 0.001$ (two-sided) the minimum size was 3 voxels (both for positive and negative weights)

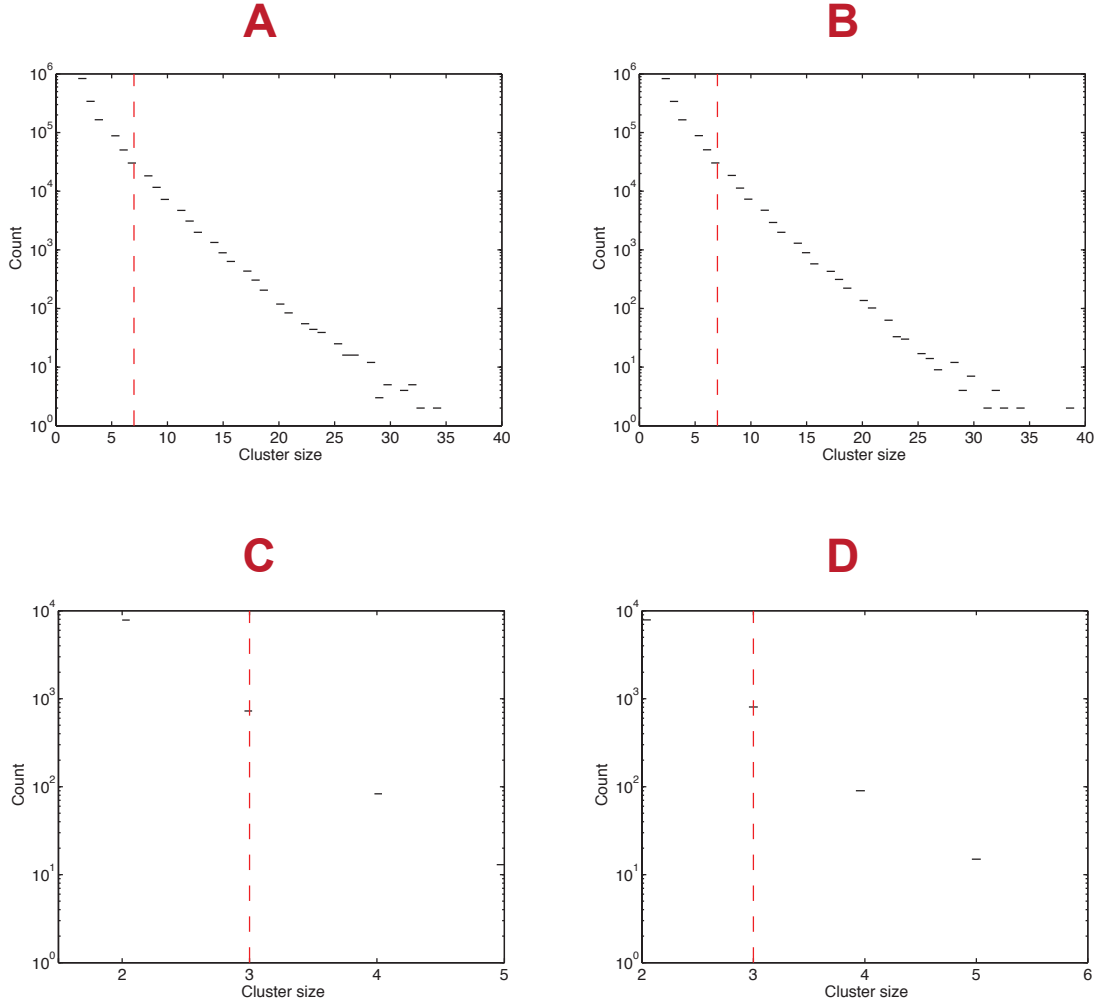


Figure 10.4: Cluster histograms for the FWM method applied to the single-subject geometric simulation data. The red line in the histogram denotes the cluster size corresponding to a cluster p -value of $p_{cl} = 0.05$ (**A**) Cluster histogram for positive weights using a voxel-wise threshold of $p_{vox} = 0.025$. The minimum cluster size for a cluster p -value $p_{cl} < 0.05$ was 7 voxels (**B**) Cluster histogram using the same voxel-wise threshold as in A ($p_{vox} = 0.025$), the minimum cluster size for $p_{cl} < 0.05$ also was 7 voxels here (**C**) Cluster histogram for positive weights and the higher voxel-wise threshold $p_{vox} = 0.0005$. The minimum cluster size for $p_{cl} < 0.05$ was 3 voxels (**D**) Cluster histogram for negative weights and $p_{vox} = 0.0005$. The minimum cluster size for $p_{cl} < 0.05$ was 3 voxels here

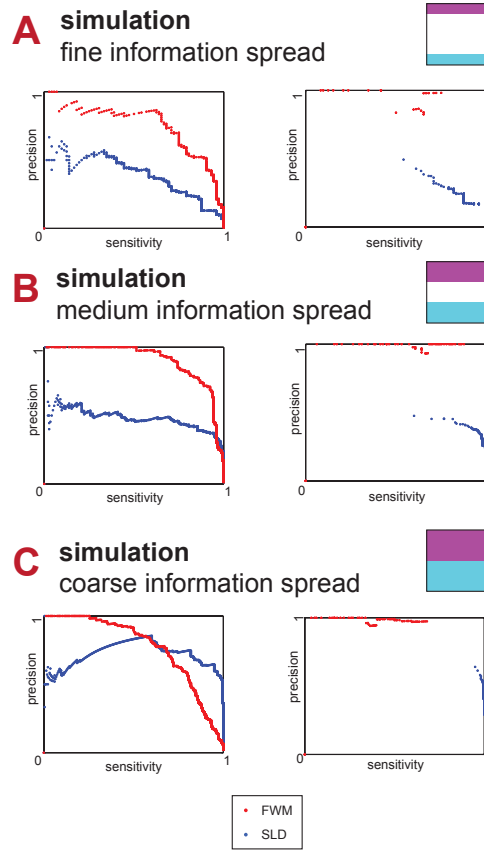


Figure 10.5: Precision/recall curves for the three different levels of information distribution of the single-subject simulation. The *precision* is the number of significant voxels inside the informative regions divided by the total number of significant voxels. The number of significantly labeled voxels inside the informative regions divided by the total volume of these regions is known as *recall* or *sensitivity*. The simulation had been cut into three equally sized areas of size $22 \times 22 \times 22$, voxels which were analyzed separately. The red dots represent the FWM method, the blue dots the SLD method. **(A)** Precision/recall curves for the leftmost third area, implementing a fine information distribution. The left box is the precision/recall curve based on uncorrected voxel p -values (derived from the permutation distribution), the right box implements the full nonparametric multiple comparison correction (where the full cluster-based analysis was carried out for a large number of p -values). **(B)** Precision/recall curves for the middle area with an intermediate distribution of information. Also here, the left box does not implement a multiple comparisons correction, while the right box is corrected using the nonparametric cluster size control. **(C)** Precision/recall curve for the rightmost area of the simulation, containing the half-cubes with a coarse distribution of information. The left box in here shows uncorrected data, while the right box is corrected for multiple comparisons with the nonparametric cluster-size based framework.

10.1.2 Influence of geometry

For analyzing the influence of geometry on the simulation, the total area was divided into three parts of equal size (i.e. $22 \times 22 \times 22$ voxels). The leftmost area contains the two half-cubes representing a fine distribution of information, the middle one an intermediate level and the rightmost one a coarse level of information distribution. The three areas had been analyzed separately from each other. The analysis was carried out with and without application of the multiple comparisons correction. If no multiple comparisons correction was used, the analysis was based on p -value maps which were derived from the permutation distribution and the original accuracy or weight maps. In both cases (uncorrected or corrected) the maps were thresholded at certain p -values, repeated by counting of the total number of significant voxels within and outside the informative regions. In case of the version with multiple comparisons correction, the entire cluster-based analysis was carried out for different levels of threshold maps.

The procedure allowed the computation of the precision (number of significant voxels within informative regions divided by total number of significant voxels) and the recall or sensitivity (which is defined as the number of significant voxels within the informative regions divided by the volume of the informative regions). In other words, the precision gives a measure of the fraction of true positives and the recall/sensitivity a measure of the fraction of informative area labeled significant. Both values are plotted against each other in Figure 10.5.

The three subdivisions of areas are displayed separately in Figure 10.5A (fine information spread), Figure 10.5B (intermediate) and Figure 10.5C (coarse information spread). The left boxes in this figure depict the uncorrected charts (without cluster-based analysis, based on permutation-derived p -values) while the right boxes implement the multiple comparisons correction (where for a large number of voxel-wise p -values the cluster-based analysis was carried out). Evidently, in the case of the uncorrected maps (left boxes), for fine and intermediate information spread the FWM method has a higher precision for any given level of sensitivity. Only in the case of coarse information spread and low thresholds (hence high sensitivity), does the SLD method return a higher precision. In general terms, the sensitivity increases for more stringent p -values (i.e. lower p -values), while the precision declines (the p -values are not displayed in this figure). For any given p -value, the FWM method and SLD method show vastly different ratios between precision and sensitivity, furthermore, this difference also depends on the underlying geometry. While the FWM method performs very well (i.e. high sensitivity and precision) for rather low p -values (e.g. $p_{vox} = 0.05$), the SLD method performs better in the regime of low p -values (e.g. $p_{vox} = 0.001$).

In the case of additional multiple comparisons using cluster-size control (the right boxes of Figure 10.5), the FWM method never achieves 100% sensitivity, i.e. does never label all informative voxels as significant. The precision, however, is extremely high in this case, indicating that the voxels labeled significant are actually almost exclusively in informative regions. The SLD approach achieves higher sensitivities in the corrected case featuring cluster size control. On the other hand the precision is very low here, in particular for the fine and intermediate information spreads. In regards to the optimal p -values, also here the FWM

p_{cl}	0.01	0.02	0.03	0.04	0.05
expected number of clusters	1	2	3	4	5
SLD number of clusters	2	2	2	4	4
FWM ⁺ number of clusters	2	2	4	4	5
FWM ⁻ number of clusters	0	0	0	1	3

Table 10.1: Results of the null simulation on single-subject level, number of expected clusters given the pre-specified type I error rate versus number of empirically found clusters. The first row depicts the 5 values for the type I error rate p_{cl} on cluster level. Given the total of 100 simulations, the number of expected clusters for each level of p_{cl} is displayed in the second row. In the third to fifth row, the number of empirically found clusters for the SLD and FWM method is displayed. For the FWM method, the statistics were carried out for positive and negative weights separately.

method performs well when using high p -values, while the SLD method performs better for lower p -values.

10.2 Single subject null simulation

A total of 100 null simulations on single-subject level were carried out. The number of permutations was set 1000 for the SLD method and 10000 for the FWM approach (due to computational limitations). The voxel-wise threshold for the SLD method was set to $p_{vox} = 0.001$, while for the FWM method the threshold was set to $p_{vox} = 0.05$ (two-sided). The type I error rate specified by the cluster level p_{cl} was varied using 5 equidistant values between 0.01 and 0.05. Hence, given the expected error rate and the number of simulations, an expectation value for the number of false positive clusters could be computed. This expectation value could be compared to the empirically found number of false positive clusters using the SLD or FWM method. The results are displayed in Table 10.1.

The results indicate that the number of empirically found clusters only marginally deviates from the number of expected clusters for value of p_{cl} . As only 100 simulations were carried out, the results should be regarded as approximation of the limit of a high number of simulations, therefore small deviations are possible. Given this consideration, it can be stated that the number of empirically found false-positive clusters lies well within the number of expected false-positive clusters for any given value of p_{cl} .

10.2.1 Influence of underlying image smoothness

In order to demonstrate how the underlying smoothness (spatial correlation) in the images is implicitly considered in the empirical cluster size histograms, I constructed 10 simulations in the same manner as the single-subject null simulation above, however, varying the Gaussian smoothing kernel FWHM using 10 equidistant values between 0 and 9 millimeters in equidistant steps, corresponding to 0 to 3 voxels (the voxel size was set to 3mm). The nonparametric statistics were computed both for the SLD method (using a voxel-wise threshold of $p_{vox} = 0.001$) and the FWM method (using a right-tailed threshold of $p_{vox} = 0.05$). Next, the cluster size

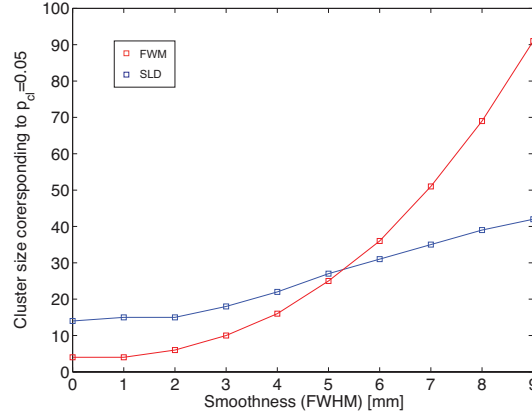


Figure 10.6: Impact of the underlying intrinsic smoothness on the cluster-size histograms. The smoothness is varied using 10 values between 0 and 9mm, corresponding to 0 to 3 voxels (with a voxel size of 3mm). The cluster size histograms for the SLD and FWM method were computed for each level of spatial correlation. This allowed a determination of the cluster size corresponding to a (uncorrected) p -value of $p_{cl} = 0.05$ (which indicates the broadness of the histogram). For both methods, this critical cluster size monotonically increases for larger values of smoothness, implying broader cluster size histograms.

corresponding to a cluster p -value $p_{cl} = 0.05$ was computed for each of the 10 cluster size histograms for the SLD and respectively for the FWM method. The results are shown in Figure 10.6. For both the SLD and FWM method, the critical cluster size (corresponding to $p_{cl} = 0.05$) monotonically depends on the smoothness of the underlying input data; more spatial correlation effectively broadens the cluster size distributions monotonically. Hence the spatial correlation between neighboring voxels is implicitly reflected in the cluster size histograms.

10.3 3T tapping synchronization experiment

The results for the 3T tapping synchronization fMRI experiment for a single-subject are displayed in Figure 10.7. The results depict the classification of a synchronization task with a *discrete* visual versus a *continuous* visual target sequence. For the analysis, the proposed nonparametric framework based on permutations and cluster size control was applied to the searchlight decoding and the feature weight mapping method. Two levels of voxel-wise thresholds were used for each method; a low one ($p_{vox} = 0.05$) and a high one ($p_{vox} = 0.001$). In the case of the FWM method, the threshold accounted for a two-sided test equivalent with two one-sided tests to each $\frac{p_{vox}}{2}$. The SLD method did not yield any significant results for the low threshold of $p_{vox} = 0.05$, therefore the threshold was decreased here to $p_{vox} = 0.01$. The threshold maps are displayed in Figure 10.8.

For the low thresholds ($p_{vox} = 0.01$), 3951 voxels were labeled significant using the SLD method. For the FWM method ($p_{vox} = 0.05$) 859 voxels were identified significant. The overlap of voxels between both methods was 684 voxels, leaving 3267 voxels identified exclusively from the SLD method and 175 voxels labeled significant solely by the FWM method. Both methods label the primary visual cortex as significant (first and second slice of Figure 10.7A/B

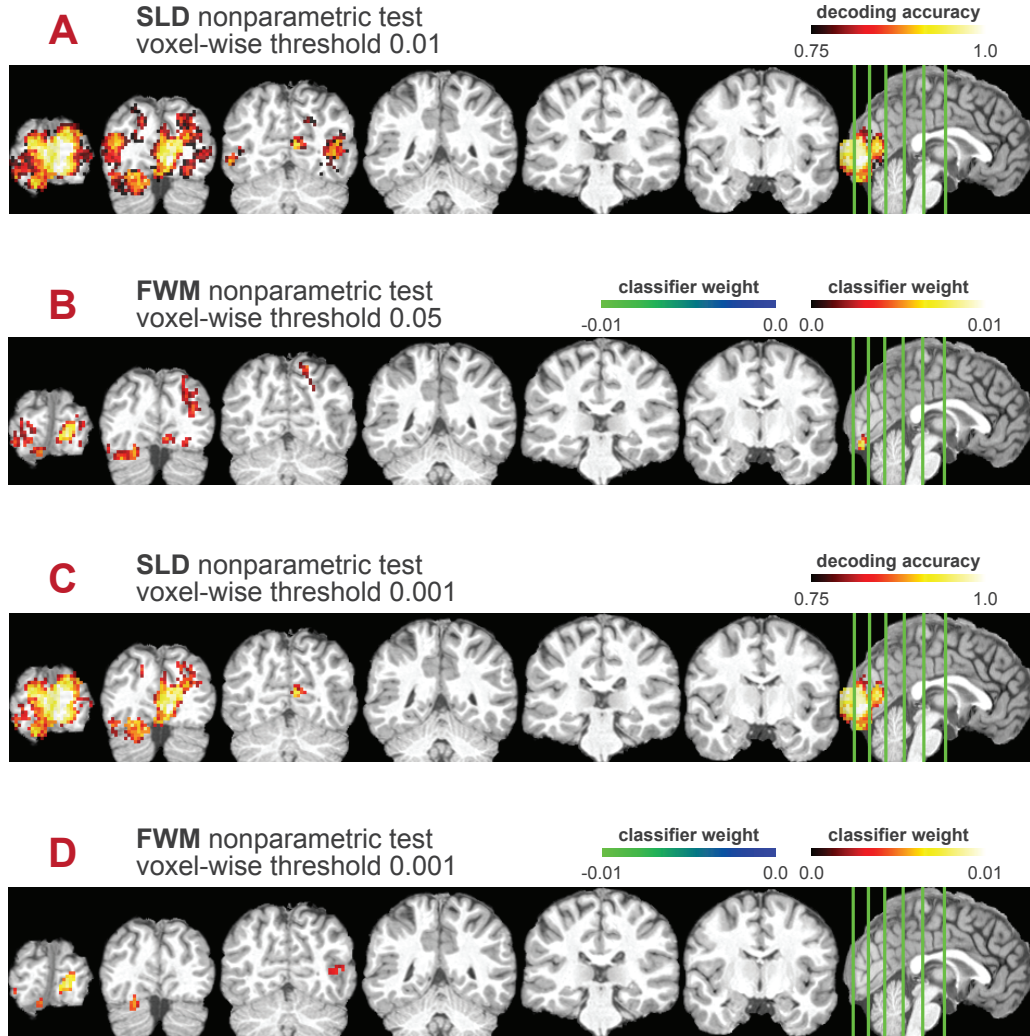


Figure 10.7: Comparison between the SLD and FWM method on single-subject level for the 3T tapping synchronization experiment. All results are corrected for multiple comparisons using the proposed nonparametric framework. **(A)** Results of the SLD method, using a voxel-wise threshold of $p_{vox} = 0.01$ (using $p_{vox} = 0.05$ no clusters were labeled significant). **(B)** Results of the FWM method using $p_{vox} = 0.05$ (two-sided). No negative weights have been labeled as significant here. **(C)** SLD results for the high voxel-wise threshold $p_{vox} = 0.001$. **(D)** FWM results for the high threshold $p_{vox} = 0.001$.

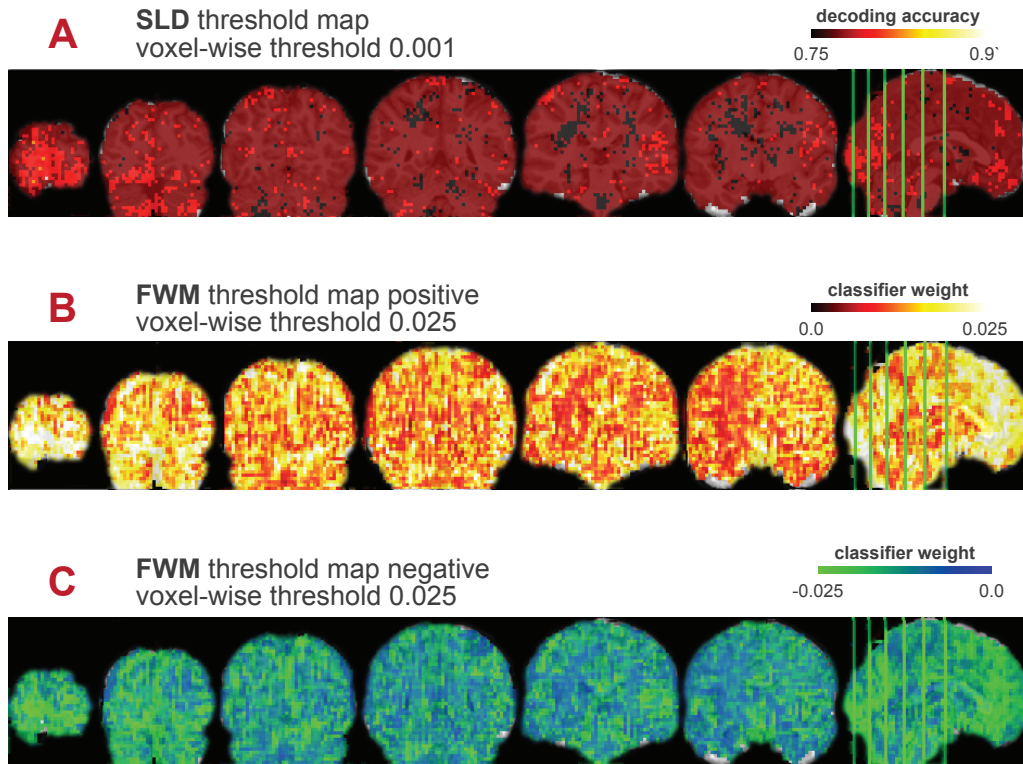


Figure 10.8: Threshold maps for the 3T tapping synchronization experiment on the single-subject level for both the SLD and FWM method. (A) SLD threshold map, depicting the accuracy level equivalent to $p_{vox} = 0.001$ (B) FWM threshold map for the positive weights and a voxel-wise threshold of $p_{vox} = 0.025$ (C) FWM threshold map depicting the negative weight level equivalent to $p_{vox} = 0.025$

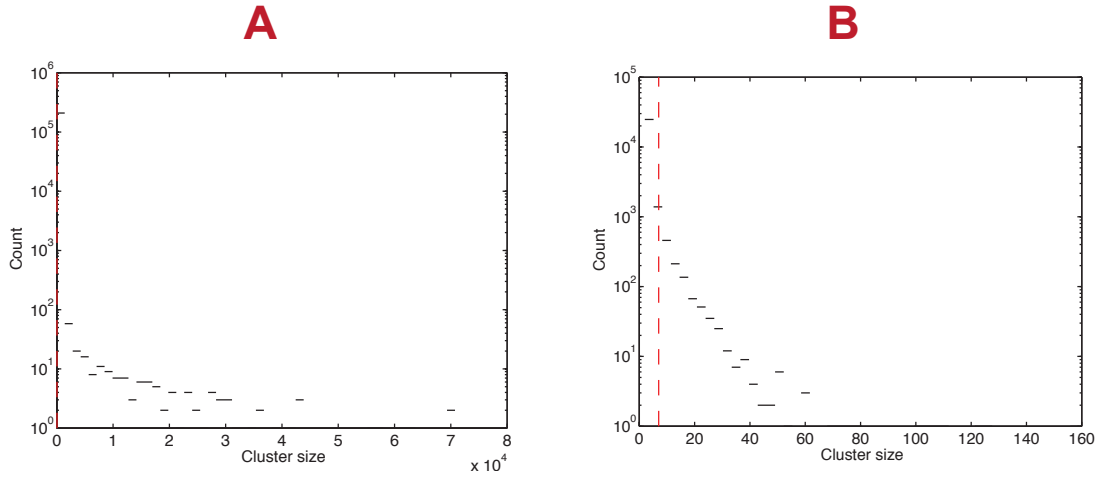


Figure 10.9: Cluster size histograms for the 3T tapping synchronization experiment using the SLD method on a single-subject level. The red line in the histograms marks the cluster size corresponding to the uncorrected cluster p -value $p_{cl} = 0.05$. **(A)** Cluster size histogram using a voxel-wise threshold of $p_{vox} = 0.01$. Clusters with a size of more than 31 voxels obtain a p -value $p_{cl} < 0.05$ **(B)** Cluster size histogram using a voxel-wise threshold was $p_{vox} = 0.001$. The cluster size corresponding to the cluster p -value of $p_{cl} = 0.05$ was 25 voxels

). Secondary visual areas and the superior parietal lobule are found bilaterally significant in the SLD method, however only on the right hemisphere for the FWM method (second slice of Figure 10.7A/B). Furthermore, the SLD method depicts the visual cortex V5, which is implicated in motion processing (third slice of Figure 10.7A). The area remains undetected in the FWM method (Figure 10.7B). Motor areas or thalamic areas are not detected on single-subject level with either method.

For the higher threshold ($p_{vox} = 0.001$), a total of 2719 voxels were labeled as significant for the SLD method. The number of significant voxels in the FWM method was 266; the overlap between both methods was 200 voxels (leaving 2519 identified exclusively by the SLD method and 66 exclusively by FWM). The SLD method labels the primary visual regions as significant (Figure 10.7C, the first two slices), furthermore secondary visual areas and the superior parietal lobule are detected. The visual area V5 is not labeled as significant at this threshold. The FWM method only labels parts of the primary visual regions as significant (Figure 10.7D at the first slice), furthermore the right visual area V5 is detected (Figure 10.7D at the third slice). Also here, neither the SLD nor the FWM method were able to identify any motor related or thalamic areas.

The cluster size histograms for the SLD method are displayed in Figure 10.9. For the lower threshold a cluster minimum size of 31 voxels was required for a cluster p -value smaller than 0.05, for the higher threshold a cluster size of at least 25 voxels was needed. The corresponding histograms of the FWM method for this data set are depicted in Figure 10.10. For the lower threshold, the minimum size for clusters to obtain a p -value smaller than 0.05 was 14 voxels (both for the positive and negative weights). In the case of the higher threshold, the minimum cluster size was 5 voxels (corresponding to a cluster p -value of $p_{cl} = 0.05$). However it should be noted that the cluster-size histograms for positive and negative weights differ slightly in their long tails; very large clusters are more probable for positive weights.

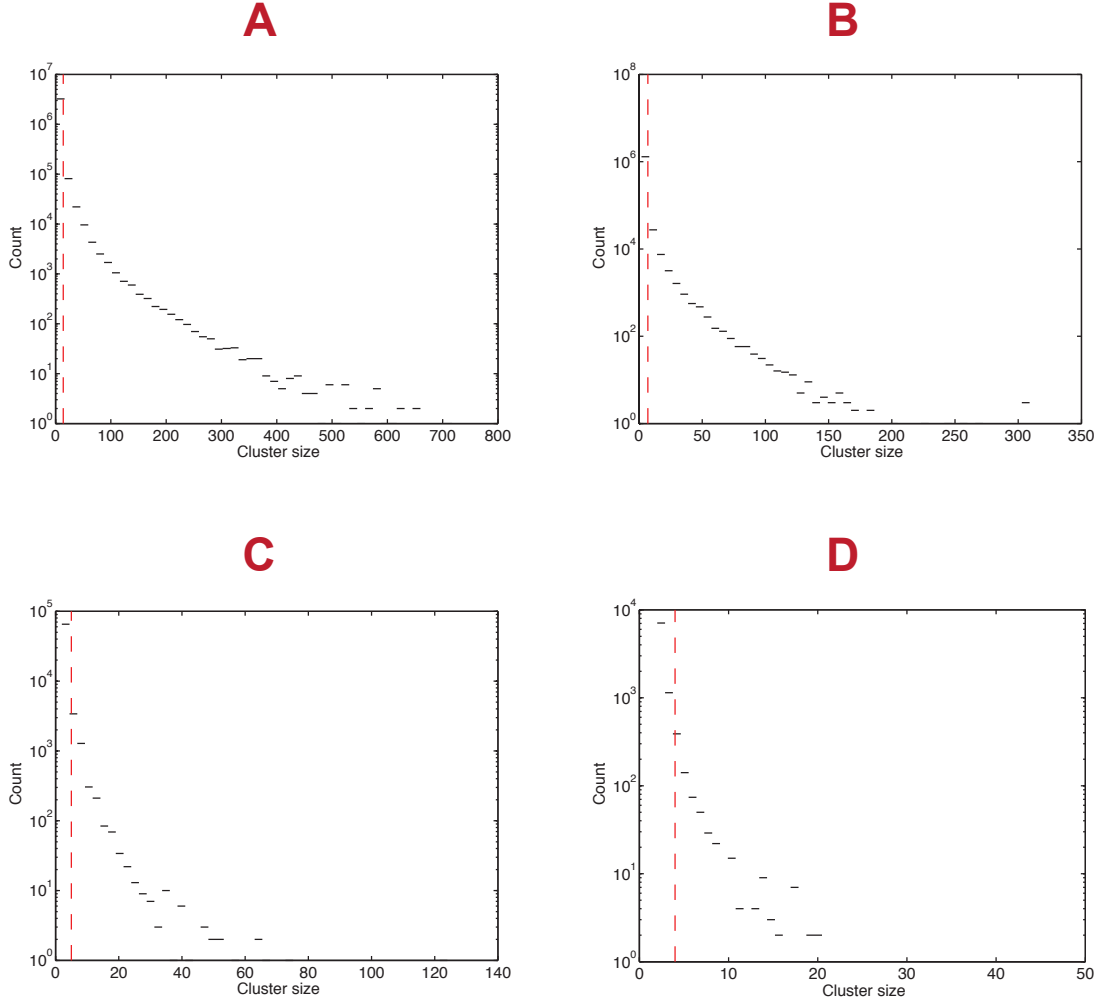


Figure 10.10: Cluster size histograms for the 3T tapping synchronization experiment using the FWM method on single-subject level. The red line in the histogram marks the cluster size corresponding to a cluster p -value of $p_{cl} = 0.05$ (**A**) Cluster histogram for positive weights using a voxel-wise threshold of $p = 0.025$. The minimum cluster size for a cluster p -value $p_{cl} < 0.05$ was 14 voxels (**B**) Cluster histogram using the same voxel-wise threshold as in A ($p = 0.025$), the minimum cluster size for $p_{cl} < 0.05$ was also 14 voxels here. (**C**) Cluster histogram for positive weights and the higher voxel-wise threshold $p_{vox} = 0.0005$. The minimum cluster size for $p_{cl} < 0.05$ was 5 voxels. (**D**) Cluster histogram for negative weights and $p_{vox} = 0.0005$. The minimum cluster size for $p_{cl} < 0.05$ was 6 voxels here.

10.4 7T finger tapping and imagination

The ultra-high resolution 7T finger tapping and imagination experiment was analyzed using the proposed nonparametric framework applied to the searchlight decoding and the feature weight mapping method. The searchlight diameter was set to 3.75mm, corresponding to a volume of 34mm^3 . Only the two conditions of rest versus tapping with touch were classified against each other. The results for a single-subject analysis are shown in Figure 10.12 (using a low voxel-wise threshold) and Figure 10.13 (using a high voxel-wise threshold). Both figures consists of three axial slices, and the slice orientation is given in Figure 10.11. The low voxel-wise threshold was set to $p_{vox} = 0.01$ in the case of SLD (as no results were significant with $p_{vox} = 0.05$), with the results shown in Figure 10.12A. The threshold for the FWM method was set to $p_{vox} = 0.05$ (two-sided), and the results are displayed in Figure 10.12B. The SLD method labels the part of the motor cortex which controls hand movements as significant. The same regions are also labeled as significant when using the FWM method. In contrast to the SLD method, however, only regions at the *surface* of the cortex are labeled significant, while the SLD method labels considerable regions as significant. As only the surface of the cortex contains grey matter and the inside of the cortex white matter, information representing the stimulus should only be contained in the surface. In other words, the SLD method labels considerable parts of uninformative regions inside the cortex as significant, which only contain white matter. Furthermore the FWM method identifies regions in the parietal and frontal cortex as significant, while these regions remain undetected for the SLD method. The total significant volume for the SLD method was 3687 voxels, and the corresponding the significant volume for the FWM method was 2065. While 865 voxels were labeled by both methods as significant, 2837 voxels were labeled as significant exclusively by the SLD method and 1215 voxels were labeled as significant only by the FWM method.

For the high voxel-wise threshold ($p_{vox} = 0.001$) the SLD method (Figure 10.13A) delineates the motor cortex in a similar fashion as when using a the lower threshold. However, additional regions in the parietal cortex are now identified as well (the same regions that had been found significant when using the FWM method with a low threshold). The (false positive) identification of white matter voxels is improved to a small degree when using the higher threshold in the SLD method. If the high voxel-wise threshold is applied to the FMW method, the results remain very sparse (Figure 10.13B); while most regions that had been identified previously with the low threshold are also found here, the significant volume shrunk considerably. The total volume labeled as significant by the SLD method was 2206 voxels, while the significant volume for the FWM method was 261 voxels. 187 voxels were labeled by both methods as significant, leaving 2019 voxels identified solely by the SLD method and 74 voxels exclusively by the FWM method at this threshold.

In Figure 10.14, the empirical cluster size histograms for the SLD method are shown. In the case of the lower voxel-wise threshold of $p_{vox} = 0.01$, a minimum size of 22 voxels was required (for a uncorrected cluster p -value smaller than 0.05). For the higher threshold $p_{vox} = 0.001$, a minimum size of 14 voxels was required. The cluster histograms for the FWM method are shown in Figure 10.15. In here, a minimum size of 8 voxels was required for the lower voxel-wise threshold of $p_{vox} = 0.05$ (two-tailed), both for the positive and negative

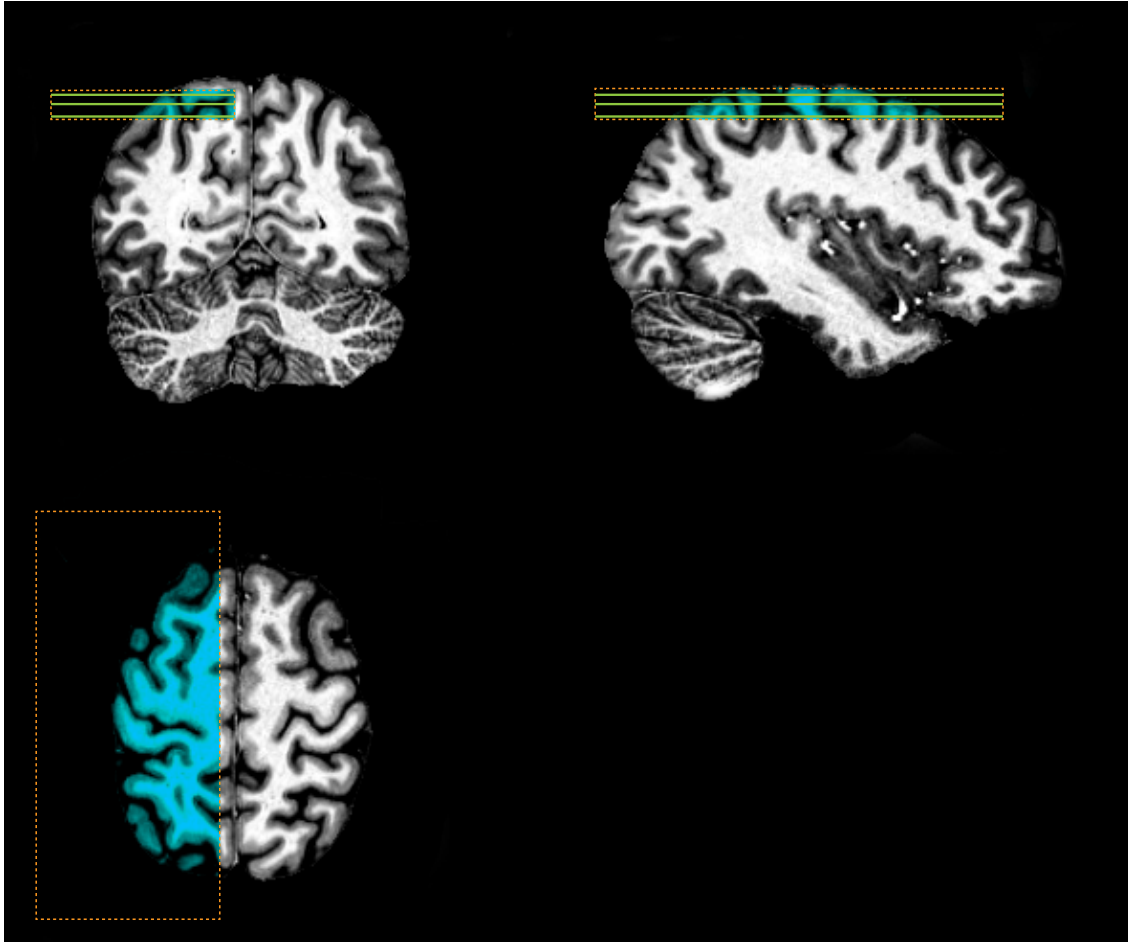


Figure 10.11: Slice orientation for the high resolution 7T finger tapping and imagination data set. The coverage is shown in light blue color and the three slice positions are depicted in green color in the coronal (upper left) and sagittal (upper right) views.

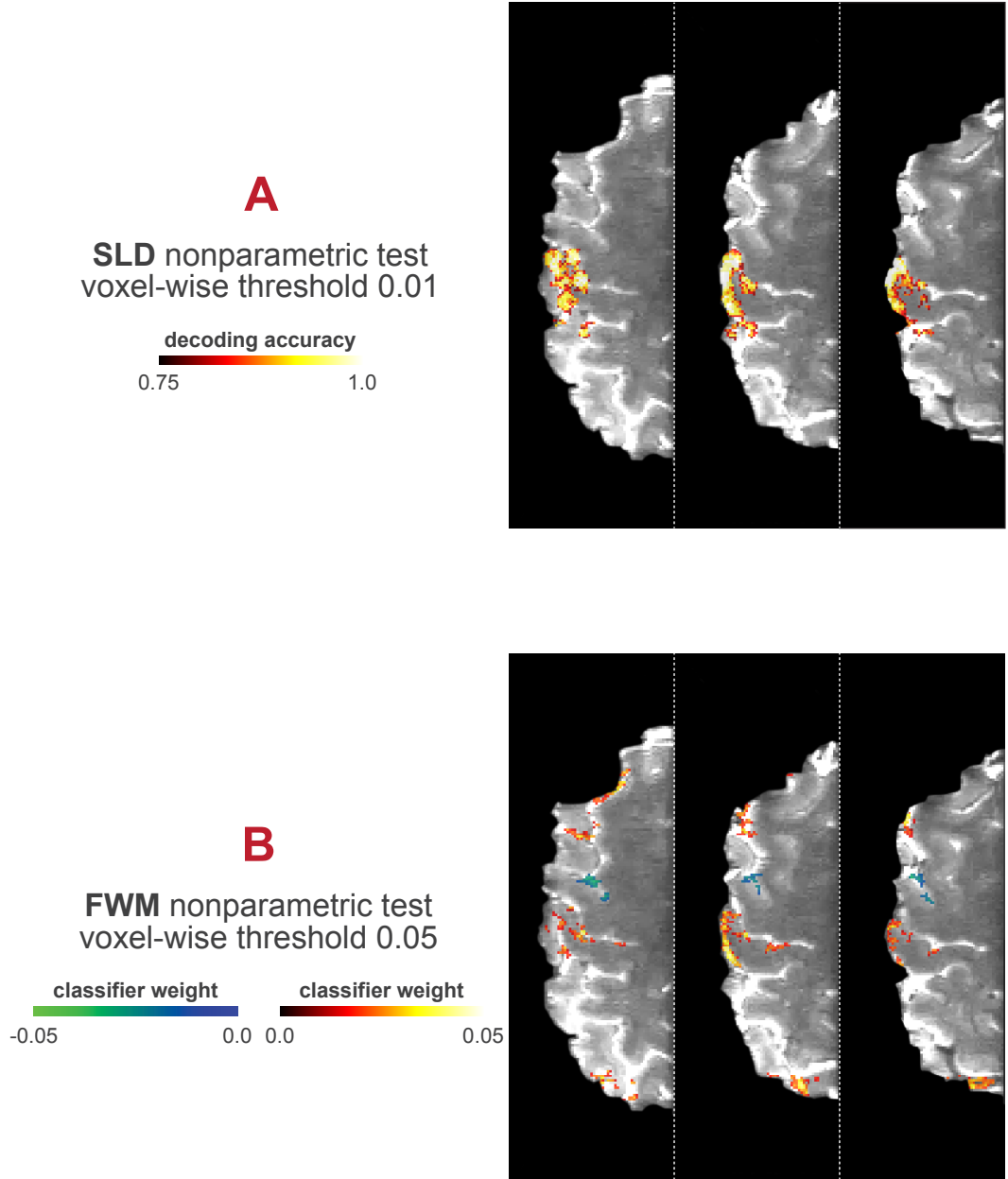


Figure 10.12: Results of the high resolution 7T finger tapping and imagination data set, classifying finger tapping with touch versus rest. The nonparametric framework proposed in this thesis had been applied to the searchlight decoding and feature weight mapping methods. This figure depicts the results for the low voxel-wise threshold. **(A)** SLD method (diameter = 3.75mm) with a voxel-wise threshold of $p_{vox} = 0.01$ (no results were returned for $p_{vox} = 0.05$). **(B)** FWM method, using a (two-sided) threshold of $p_{vox} = 0.05$.

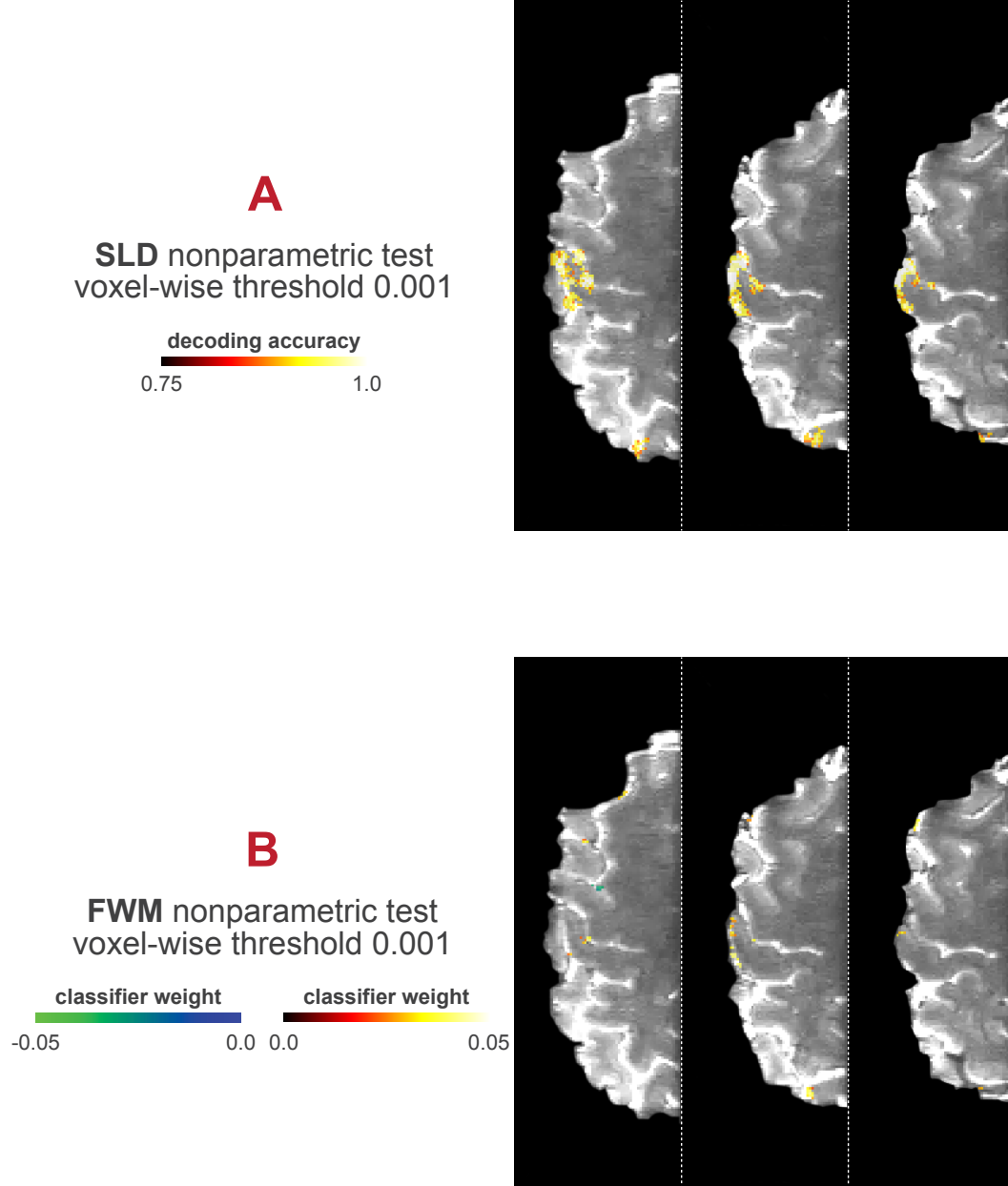


Figure 10.13: Results of the high resolution 7T finger tapping and imagination data set, classifying between finger tapping with touch versus rest. The nonparametric framework proposed in this thesis had been applied to the searchlight decoding and feature weight mapping method. The current Figure shows the results for the high voxel-wise threshold ($p_{vox} = 0.001$). **(A)** SLD method (diameter = 3.75mm) **(B)** FWM method, using a (two-sided) threshold of $p_{vox} = 0.001$.

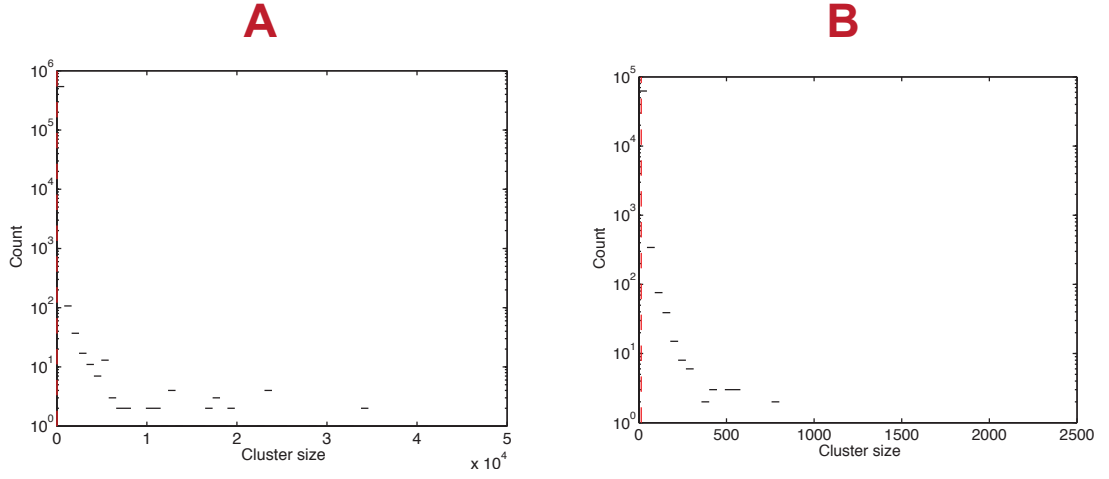


Figure 10.14: Cluster histograms for the SLD method, which was applied to the 7T finger tapping and imagination data set on a single-subject level. The red line in the histograms denotes the cluster size corresponding to the uncorrected cluster p -value of $p_{cl} = 0.05$. **(A)** Cluster-size histogram using a voxel-wise threshold of $p_{vox} = 0.01$. Clusters with a size larger than 22 voxels obtain a p -value $p_{cl} < 0.05$ **(B)** Cluster-size histogram using a voxel-wise threshold of $p_{vox} = 0.001$. The minimum cluster size corresponding to the cluster p -value $p_{cl} = 0.05$ was 14 voxels.

weights. In the case of the higher voxel-wise threshold $p_{vox} = 0.001$, the minimum cluster size corresponding to a cluster p -value of smaller than 0.05 was 3 voxels both for the positive as well as negative weights.

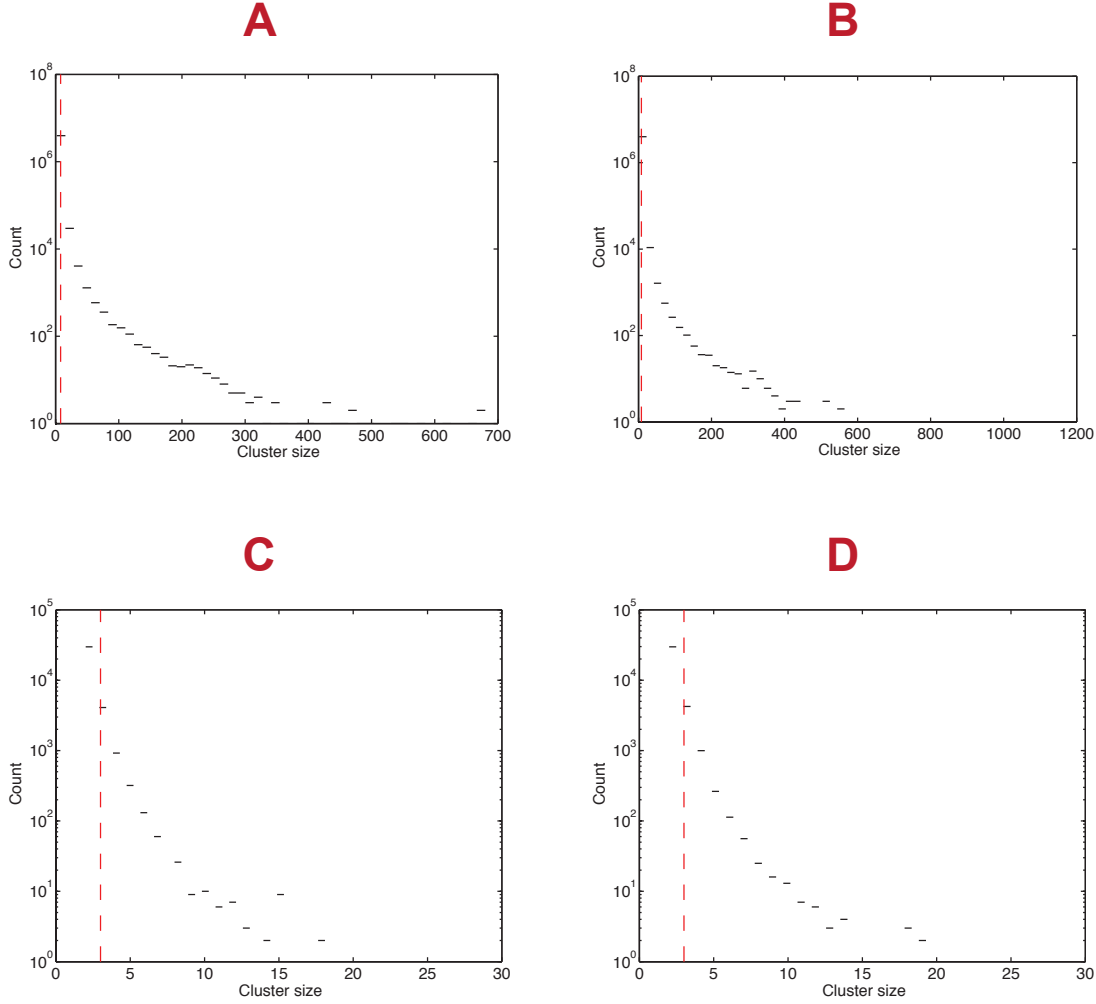


Figure 10.15: Cluster histograms for the FWM method applied to the 7T finger tapping and imagination data set. The red line in the histogram marks the cluster size corresponding to a cluster p -value of $p_{cl} = 0.05$. **(A)** Cluster histogram for positive weights using a voxel-wise threshold of $p = 0.025$. The minimum cluster size for a cluster p -value of $p_{cl} < 0.05$ was 8 voxels. **(B)** Cluster histogram using the same voxel-wise threshold as in *A* ($p = 0.025$), the minimum cluster size for $p_{cl} < 0.05$ was also 8 voxels here. **(C)** Cluster histogram for positive weights and the higher voxel-wise threshold $p_{vox} = 0.0005$. The minimum cluster size for $p_{cl} < 0.05$ was 3 voxels. **(D)** Cluster histogram for negative weights and $p_{vox} = 0.0005$. The minimum cluster size for $p_{cl} < 0.05$ was also 3 voxels here.

Chapter 11

Group analysis results

11.1 Group simulation 5cubes

The aim of this simulation was to emulate a *virtual group* of subjects which are in two distinct “brain”-states. Most crucially, the simulation allowed a precise determination of the local information content. This made a quantitative comparison possible between the proposed non-parametric framework and T-based methods. The comparison was carried out for both the searchlight decoding and for the feature weight mapping method. Furthermore, the simulation allowed a direct comparison between the searchlight method and feature weight method itself (using the nonparametric framework). It should be noted that since in this simulation only a positive class offset was added to the data points of one class, only a *one-sided* threshold was applied for the FWM method.

11.1.1 Nonparametric vs parametric

In the following, the proposed nonparametric framework for group level analysis is compared to the commonly applied T-based method for statistical inference. The comparisons are carried out separately for the SLD and FWM method.

11.1.1.1 Searchlight decoding

Comparison of nonparametric vs parametric A detailed comparison between the proposed method and T-based statistics is found in Figure 11.1. Here, one slice of the simulated data is shown which includes the five cubes where information had been added. The arrangement of the cubes is displayed in Figure 11.1A. The white regions indicate the five informative cubes and the black regions represent the noise background. The raw group accuracy map, which is the average of the 12 single-subject accuracy maps, is depicted in Figure 11.1B. Upon visual inspection, it is possible to locate the five informative regions, while in some other areas (especially on the right side) the noise background appears with a very similar structure. Figure 11.1C shows the

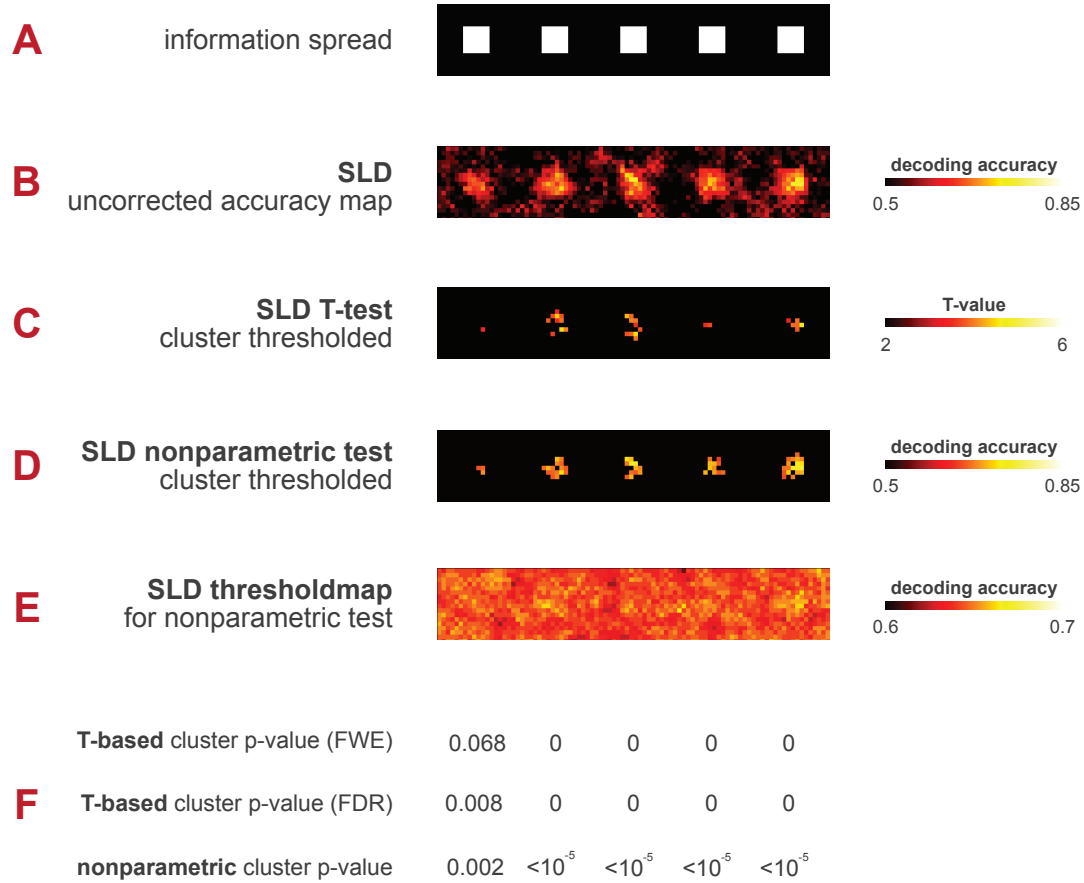


Figure 11.1: Group simulation 5cubes data set analyzed by the SLD method comparing the parametric T-based versus the proposed nonparametric framework. **(A)** Information spread of the raw data (which served as the input for the classifier). Note the five cubes where information was stored and the variation of information content: the first cube from the left had the smallest information content, progressing to the last on the right where the information was at the maximum. **(B)** Mean decoding accuracy map over all 12 virtual subjects. **(C)** T-test on the accuracy maps, corrected for multiple testing using Gaussian random fields and FDR cluster thresholding (SPM8) and a voxel-wise threshold of $p_{vox} = 0.001$. **(D)** Results of the new nonparametric method based on permutation and Monte-Carlo resampling methods implementing the multiple comparisons correction. The voxel-wise threshold $p_{vox} = 0.001$ was set here to the same value as for the T-based method. **(E)** Threshold map for the cluster search algorithm, depicting the accuracy corresponding to $p_{vox} = 0.001$. The map displays an inhomogeneity of the local chance distribution. **(F)** Table with the cluster p -values for each of the five cubes (represented as columns) using different statistical measures (represented as rows). The first two rows display T-based methods, the last one indicates the new nonparametric method.

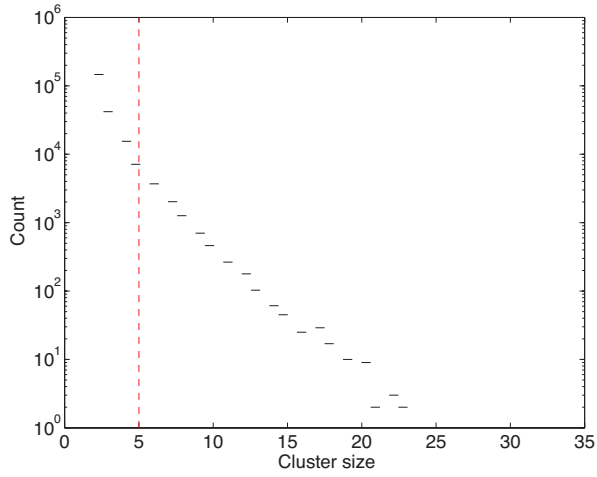


Figure 11.2: Nonparametric cluster-size histogram from the group simulation 5cubes data set based on searchlight decoding and a voxel-wise threshold $p_{vox} = 0.001$. The red line marks the (uncorrected) $p_{cl} = 0.05$ percentile of the cluster size distribution (cluster size = 5 voxels).

standard T-test against chance level carried out in SPM8. The map was thresholded at a voxel-level of $p_{vox} = 0.001$ and corrected by Gaussian random fields methods and FDR-methods. The minimum cluster p -level was set to $p_{cl} = 0.05$. On the other hand, Figure 11.1D demonstrates the new nonparametric cluster-size control, including multiple comparisons correction and using the same voxel-level threshold of $p_{vox} = 0.001$. The map here shows the supra-threshold voxel-wise accuracy values. The (nonparametric) threshold map for a voxel-wise $p_{vox} = 0.001$ level is displayed in Figure 11.1E, revealing the spatial inhomogeneity found in the width of the chance distribution. Figure 11.1F shows the detailed statistics for each cluster for both the T-based and the nonparametric approaches. Out of the five informative regions in our data simulation, four (FWE controlled) or all five (FDR controlled) were revealed using SPM8 and T-tests. When the proposed method was used, all five informative regions could be decoded. The total informative area principally contained $5 \cdot 6^3 = 1080$ voxels in total. The standard SPM8 method with FDR correction on cluster level labeled approximately 12% of this volume as significant (134 voxels), while the nonparametric approach determined that approximately 24% of the informative volume was significant (258 voxels). This represents an increase of almost 100% in terms of sensitivity.

The cluster distribution of the 10^5 chance group accuracy maps from the nonparametric framework is shown in Figure 11.2. The red line indicates the cluster size corresponding to the $p_{cl} = 0.05$ level, where the right tail area of the (normalized) cluster size distribution is smaller than 0.05. For this p -value (or smaller), clusters need to have an extent of five voxels or more. Notably, this is the uncorrected threshold displayed for illustrative purposes here; in the framework, a step-down FDR procedure had been applied to the uncorrected p -values and hence a corrected threshold was set.

Influence of the initial voxel-wise threshold The voxel threshold p_{vox} was varied (i.e. the threshold for a voxel to be counted as belonging to a cluster), in order to investigate its impact on the cluster statistics. A total of four different voxel thresholds were used ($p_1 = 0.05$, $p_2 = 0.01$, $p_3 = 0.005$, $p_4 = 0.001$). For each threshold both the nonparametric method and

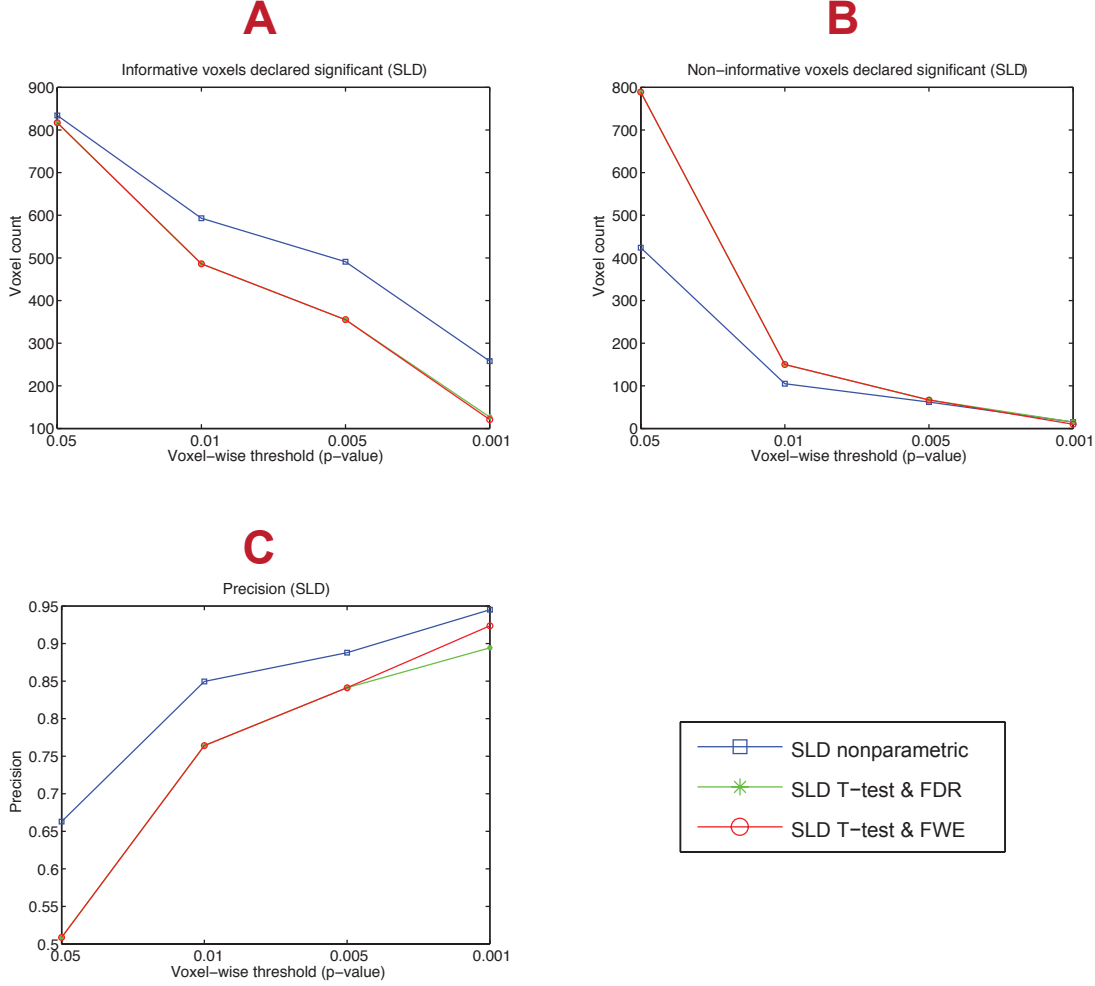


Figure 11.3: Influence of the initial voxel threshold p_{vox} for the SLD method in the group simulation 5cubes. **(A)** Number of voxels declared significant, which were located within the informative regions. With higher thresholds (i.e., smaller values for p_{vox}), the number of voxels declared significant decreases. For all values of p_{vox} , the proposed method declared more voxels significant as compared to the T-test based frameworks. **(B)** Number of voxels declared significant in non-informative regions. Also here, for higher thresholds (smaller p -values) less voxels are declared significant. For the three smallest thresholds (p_1 , p_2 and p_3) the proposed nonparametric method declares a smaller fraction of voxels significant in the non-informative regions. For the highest threshold (p_4) the number of significant non-informative voxels is comparable between the nonparametric and T-based methods. **(C)** Precision, i.e. ratio of voxels labeled significant within informative regions and the total number of voxels labeled significant. Throughout all thresholds (p_1 to p_4) the nonparametric method has a higher precision as compared to T-based approaches.

the standard T-test approaches were computed. Next, the subset of voxels that surpassed the multiple comparisons corrections were further investigated in regards to whether they were located inside or outside the informative regions (the 5 cubes). The results are displayed in Figure 11.3. It should be mentioned that I apply a rather strict interpretation of false positivity in the sense that if the center voxel of the searchlight is outside the informative region and the searchlight is declared significant, the voxel is counted as false positive (even if a part of the searchlight volume actually is within an informative region). For the lowest threshold (p_1), the amount of significant voxels in informative regions is comparable for all three implementations (nonparametric, T-test and FDR, T-test and FWE). However, the number of voxels declared significant in non-informative regions by T-tests for this threshold is very high, and about the same as the number of voxels declared significant within the informative regions. For every higher threshold (p_2 to p_4), the total number of voxels declared significant that are located in informative regions is larger when the new proposed nonparametric method is used (see Figure 11.3A). The number of significant voxels outside of informative regions is lower for the nonparametric technique for all thresholds (see Figure 11.3B). The precision (ratio between the volume labeled significant within the informative regions and the total volume labeled significant) is shown in Figure 11.3C. For the all thresholds (p_1 to p_4), the nonparametric method has a higher precision here.

Influence of the searchlight diameter I varied the searchlight diameter over five values within the group simulation 5cubes: three, five, seven, nine and eleven voxels were used. Both the proposed nonparametric method and standard T-based methods were analyzed with the different diameters. For this, I calculated the number of voxels declared significant inside and outside the informative regions. Note that as the searchlight diameter increases, this rather strict measure *penalizes* the searchlight method *per se*, as the center voxel of a searchlight may be outside the informative region but a (increasingly larger) part of its content actually stems from the informative regions. In other words, when larger searchlights are employed, the likelihood of mapping accuracy values outside the informative regions increases, because a sufficient fraction of the searchlight might lie inside informative regions while the center does not. The results are depicted in Figure 11.4. Throughout all searchlight diameters, the new nonparametric method defines a larger number of voxels as significant. For small searchlight volumes (three and five voxels) a volumetric gain of more than 100% can be achieved (as compared to T-based methods), while for larger diameters the gain is about 50–70% (Figure 11.4A). The number of voxels declared significant outside the informative cubes, however, is larger for the nonparametric method, especially for larger searchlight diameters (Figure 11.4B). The precision (ratio of significant volume within informative regions and total significant volume) is displayed in Figure 11.4C and is higher for smaller diameters when using the nonparametric method. For large searchlight diameters, however, the T-based method exhibits a higher value for the precision.

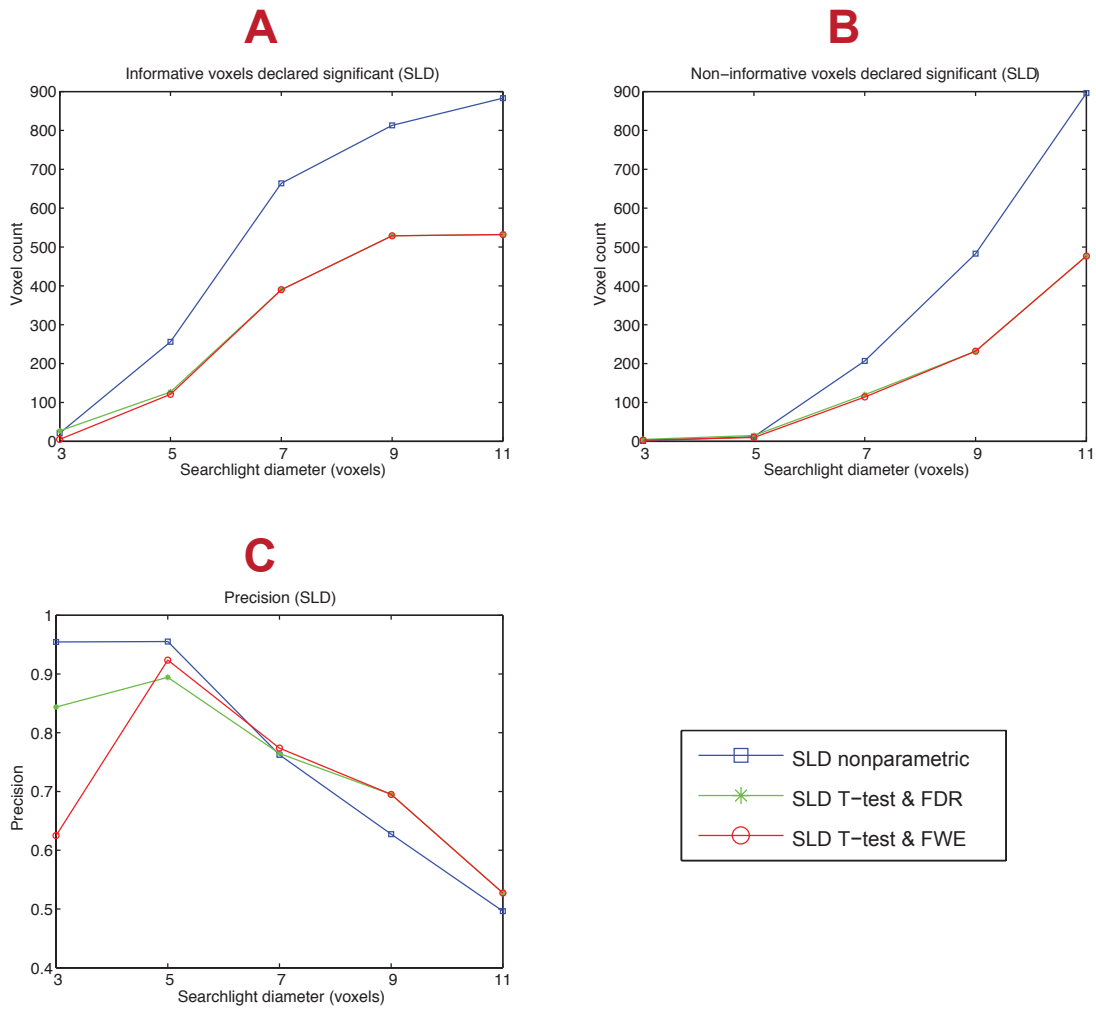


Figure 11.4: Influence of searchlight diameter in SLD maps from the group simulation 5cubes. The number of voxels within and outside the informative regions for five different searchlight diameters was computed. **(A)** Number of voxels inside the informative cubes declared significant. **(B)** Number of voxels outside the informative cubes declared significant **(C)** Precision, i.e. ratio of voxels labeled significant within the informative region and the total volume labeled significant

11.1.1.2 Feature weight mapping

In the following, the FWM method was applied to the group simulation 5cubes data. It should be noted that since there was only a *positive* offset in one group in one condition of this simulation, only *one* threshold map was computed. As before, I begin with a comparison between the T-based method and the proposed nonparametric framework.

Comparison of nonparametric vs parametric An overview over the results can be found in Figure 11.5. The same slice as in the analogous SLD figure (Figure 11.1) is displayed here. The white cubes in Figure 11.5A also display the informative regions in white color. The mean of the feature weights (of the 12 virtual single subjects) is displayed in Figure 11.5B. The result of the one-tailed T-test including a multiple comparisons correction (Gaussian random fields methods and FDR on the cluster p -values) is shown in Figure 11.5C. The voxel-wise threshold was set to $p_{vox} = 0.05$. Figure 11.5D displays the result of the proposed nonparametric statistical framework. The voxel-wise threshold $p_{vox} = 0.05$ was identical to the T-based methods. The corresponding nonparametric threshold map is depicted in Figure 11.5E, which marks the feature weight corresponding to a voxel-wise p -value of $p_{vox} = 0.05$. The threshold map shows inhomogeneities in the widths of the underlying voxel-wise null distributions. The cluster p -values for each of the five clusters are shown in Figure 11.5F. Using FWM, regardless of the type of statistic (nonparametric or parametric using FDR or FWE on cluster level), all informative cubes could be identified.

Of the total informative area ($5 \cdot 6^3 = 1080$ voxels), about 55% was labeled significant with the T-based methods (591 voxels). With the nonparametric method, 54% of the informative volume was labeled significant (580 voxels). Hence, in terms of sensitivity, the methods are comparable. However, the volume labeled significant *outside* of the informative regions was 113 voxel in T-based methods, as compared to 34 voxels for the proposed nonparametric framework.

The cluster-size distribution from the 10^5 resampled chance feature weight maps is displayed in Figure 11.6. The (uncorrected) cluster-size threshold is indicated with the red line; in here the right-tailed area of the normalized cluster size distribution is smaller than 0.05. The corresponding cluster size is 2 voxels or larger. Importantly, this is the uncorrected cluster threshold; in the framework a step-down FDR procedure corrects the p -values of the identified clusters.

Influence of the initial voxel-level threshold As the result of the statistical frameworks highly depend on the initial voxel-wise threshold p_{vox} , the impact of choice of this parameter on the cluster statistics was investigated. In here, four different p -values were used ($p_1 = 0.05$, $p_2 = 0.01$, $p_3 = 0.005$, $p_4 = 0.001$). For each threshold both the nonparametric method and the standard T-test approaches were computed.

The subsets of voxels surpassing the multiple comparisons correction for each of the four threshold values were investigated in regards to their spatial distribution. For this, the

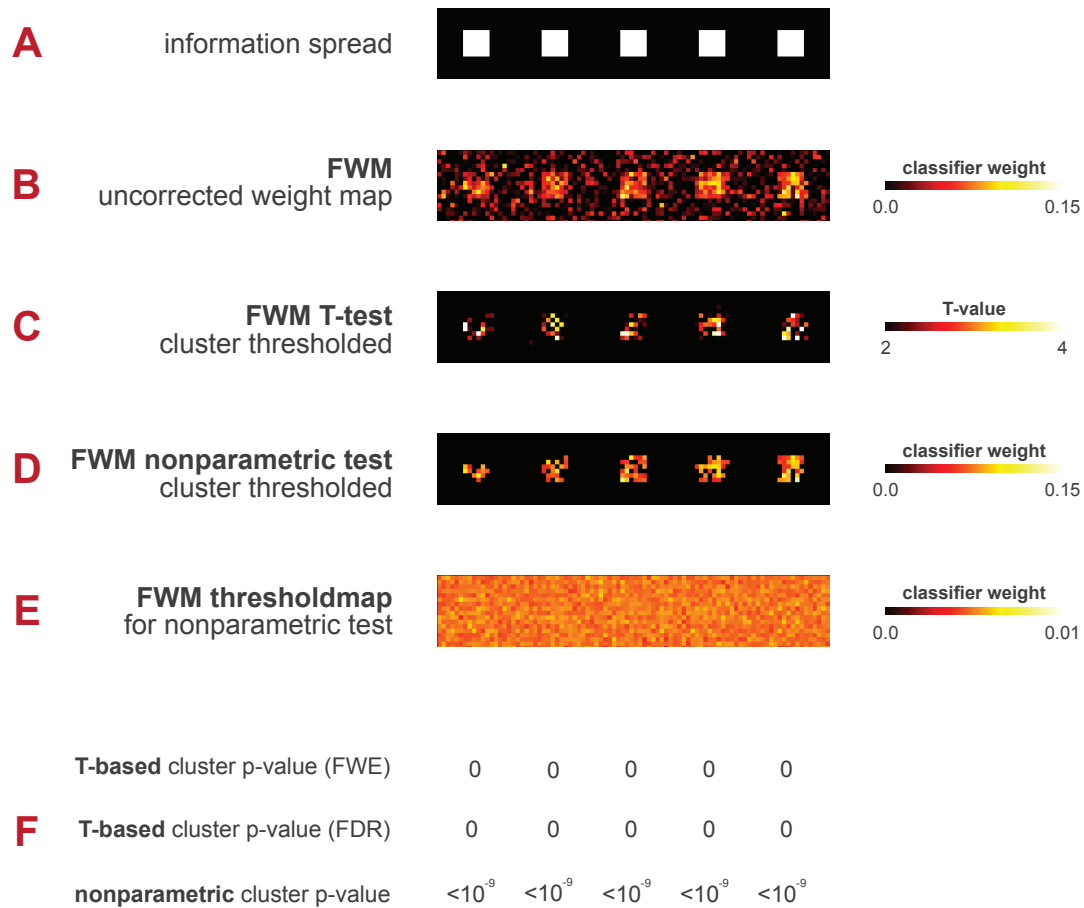


Figure 11.5: Group simulation 5cubes data set analyzed by the feature weight mapping method using the proposed nonparametric inference and T-based methods. **(A)** Information spread of the raw data, class information had been present within the white cubes. The first cube from the left contained the smallest information content and the last on the right had the maximum amount of information (realized by an additive offset in one data class). **(B)** Mean feature weights over all 12 virtual subjects. **(C)** T-test on the feature weight maps, corrected for multiple testing using Gaussian random fields and FDR cluster thresholding (SPM8). The voxel-wise threshold was set to $p_{vox} = 0.05$. **(D)** Thresholded feature weight maps from the new nonparametric method based on permutation and Monte-Carlo resampling methods implementing cluster size correction, at the same voxel-wise threshold of $p_{vox} = 0.05$. **(E)** Threshold map for the cluster search algorithm, indicating the inhomogeneity of the threshold weights at a voxel-wise p -level of $p_{vox} = 0.05$. **(F)** Table with the cluster p -values for each of the five cubes (represented as columns) using different statistical measures (represented as rows). The first two rows display T-based methods, the last one indicates the new nonparametric method.

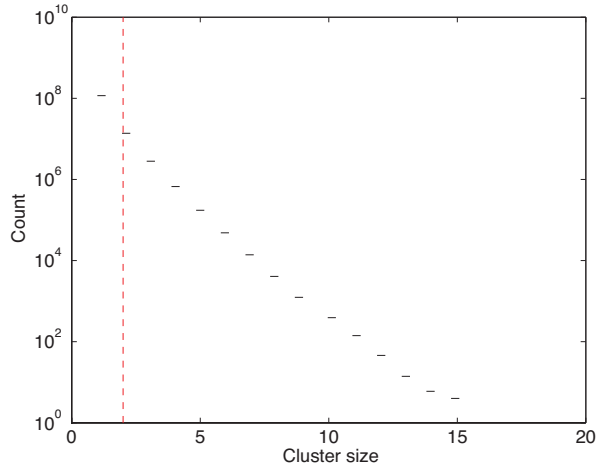


Figure 11.6: Cluster size histogram from the group simulation 5cubes data set for nonparametric analysis based on feature weight mapping and a voxel-wise threshold of $p_{vox} = 0.05$. The red line marks the $p_{cl} = 0.05$ (uncorrected) percentile of the cluster size distribution (cluster size = 2 voxels).

number of supra-threshold voxels *inside* the informative regions (the 5 cubes) was computed as well as the number of supra-threshold voxels *outside*. As already done for the simulation analyzed by the SLD method (see Section 11.1.1.1 on page 101), voxels labeled outside the informative regions are strictly regarded as false positives. The results for the four voxel-wise thresholds are shown in Figure 11.7. In the case of the lowest threshold p_1 , the number of true positives is comparable for parametric and nonparametric methods. For the three higher thresholds (p_2 to p_4), the nonparametric method surpasses the T-based approach in terms of the number of correctly identified informative voxels (see Figure 11.7A). On the other hand, the number of voxels that were falsely declared as significant, i.e. voxels residing outside of the informative region, was drastically larger for the T-based methods in the case of the lowest threshold p_1 . For the remaining three higher thresholds, both parametric and nonparametric methods performed comparably (see Figure 11.7B). The ratio between the number of voxels labeled as significant within the informative regions and the total number of voxels determined significant is depicted in Figure 11.7C. Only for the lowest threshold p_1 , does the ratio show a noteworthy difference, in here the ratio of the nonparametric method is considerably higher.

11.1.2 Searchlight decoding vs feature weight mapping

Qualitative comparison For the comparison, the nonparametric framework is applied both to the searchlight decoding and to the feature weight method, hence making it possible to compare the outcome of both methods directly. On a qualitative level, the results of the comparison can be found in Figure 11.8. In here, two voxel-wise p -values were selected ($p_1 = 0.05$ and $p_2 = 0.001$). The results for the first threshold are depicted in Figure 11.8B and 11.8C. The SLD method is able to capture a large part of the informative areas, while the results of the FWM method are much more sparse. On the other hand, the SLD method labels more volume outside the informative cubes as significant. For the smaller threshold p_2 (results shown in Figure 11.8C and 11.8D), the SLD method clearly outperforms the FWM method, as the latter is only able to identify two of the informative cubes to a partial degree (as opposed to all five cubes for the SLD method).

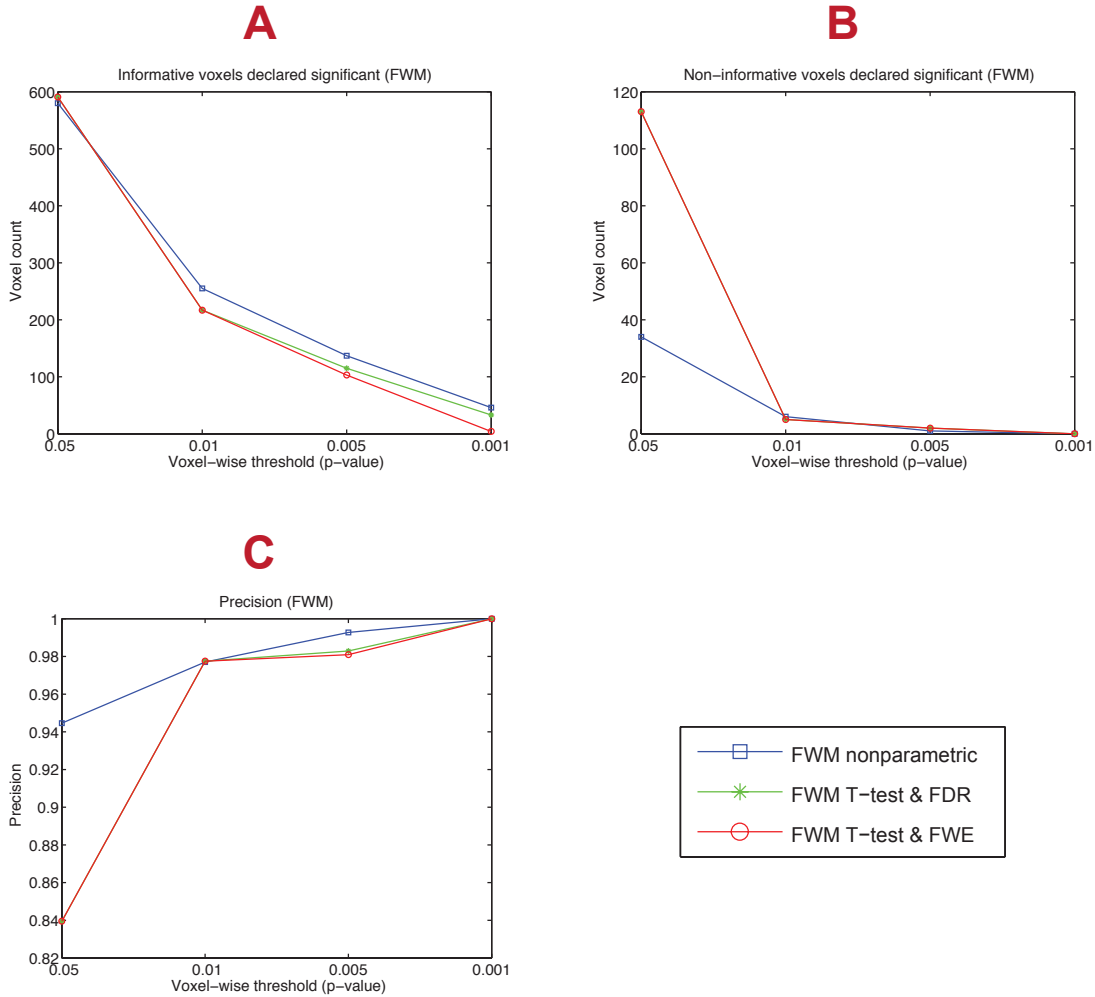


Figure 11.7: Group simulation 5cubes, influence of the initial voxel threshold p_{voxel} in the case of feature weight maps. **(A)** Number of voxels declared significant within the informative areas. With higher thresholds (i.e. smaller initial voxel p -values), the number of voxels declared significant decreases. For the three lowest p -values, the proposed method declares more voxels significant as compared to the T-test based frameworks. **(B)** Number of voxels declared significant in non-informative regions (false positives). For higher thresholds (smaller p -values) the number of voxels labeled significant is smaller. The number of falsely significant voxels is considerably lower for the nonparametric method at the lowest threshold p_1 . The results of the higher thresholds are comparable. **(C)** The precision of the nonparametric method is considerably higher for the lowest threshold p_1 , while the precision values are comparable for the higher ones.

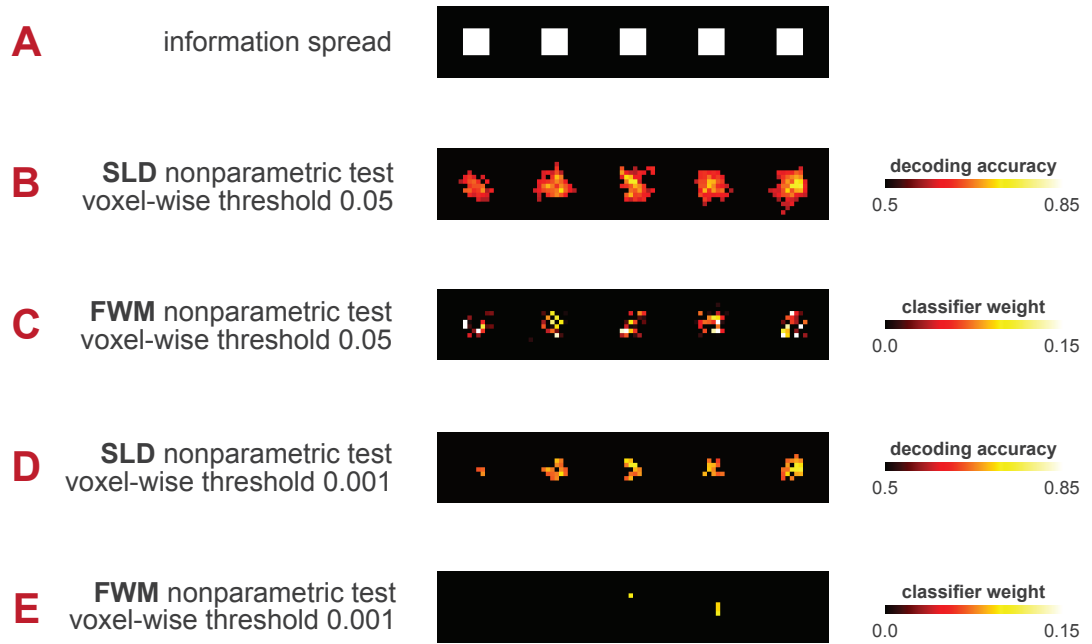


Figure 11.8: Comparison between searchlight decoding and feature weight mapping in the group simulation 5cubes data set. Both approaches have been corrected for multiple comparisons using the nonparametric framework, implementing permutation tests and Monte-Carlo resampling for the group analysis. Two different values for the initial voxel-wise threshold were used ($p_{vox} = 0.05$ and $p_{vox} = 0.001$). **(A)** Information spread of the raw data, within the white cubes class information was present (in the form of an additive offset in one data class). **(B)** Searchlight decoding map with a voxel-wise threshold of $p_{vox} = 0.05$. **(C)** Feature weight mapping, using the same threshold of $p_{vox} = 0.05$. **(D)** Searchlight decoding map implementing the lower threshold of $p_{vox} = 0.001$ **(E)** The corresponding feature weight map with $p_{vox} = 0.001$.

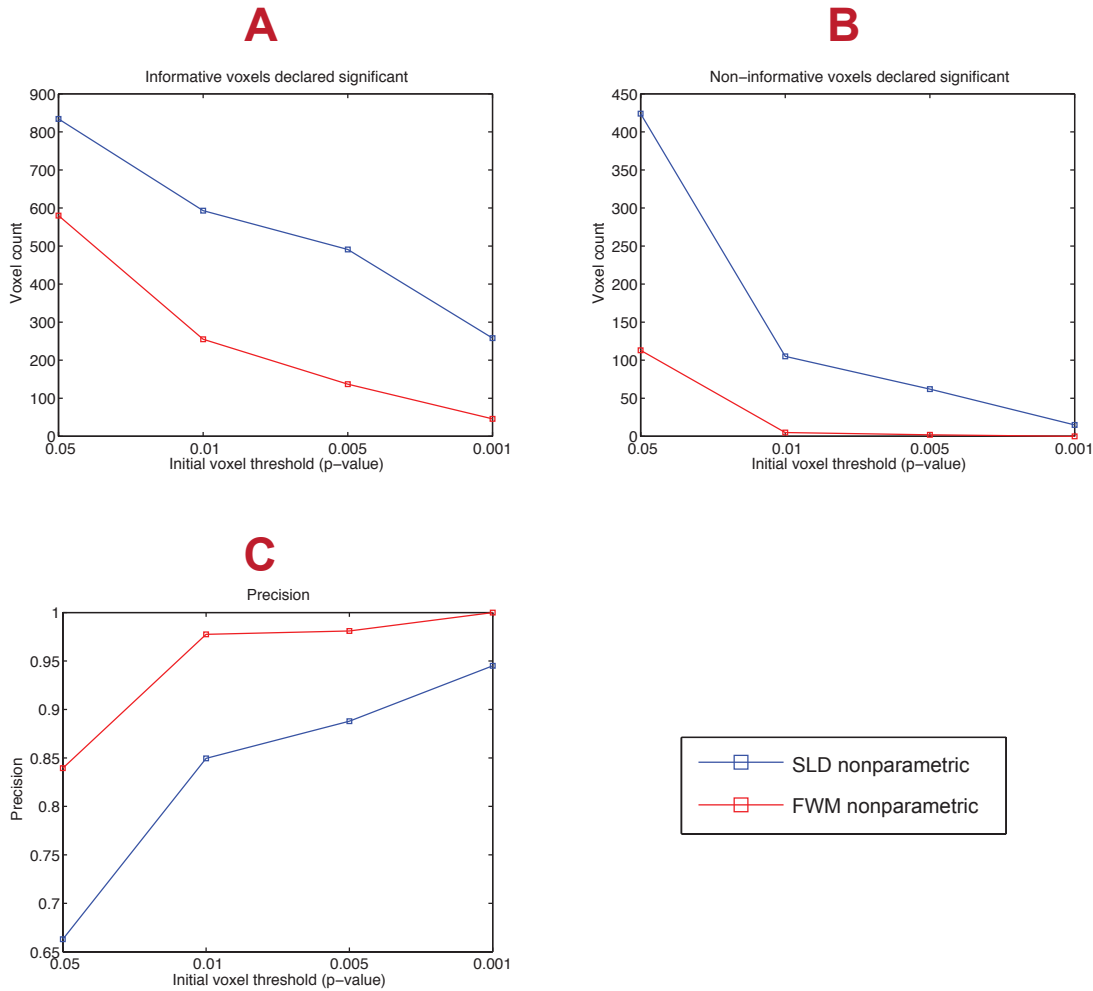


Figure 11.9: Influence of the voxel-wise threshold on the SLD and FWM method in the group simulation 5cubes, shown in blue and red colors respectively. Both results are computed with the nonparametric framework. **(A)** Number of voxels inside the informative cubes declared significant. **(B)** Number of voxels outside the informative cubes declared significant. **(C)** Precision, i.e. ratio of voxels labeled significant within the informative region and the total volume labeled significant.

Influence of voxel-wise threshold In the following, the impact of the voxel-wise threshold is investigated on a quantitative basis. As before, the number of significant voxels both inside and outside the informative regions is computed, furthermore allowing a calculation of the precision. This is done for each of the four voxel-wise thresholds $p_1 = 0.05$, $p_2 = 0.01$, $p_3 = 0.005$ and $p_4 = 0.001$.

As was already visible in the qualitative comparison above, the number of significant informative voxels is considerably larger for the SLD method (Figure 11.9A). This is true for all of the four thresholds. On the other hand, the number of voxels labeled significant outside of the informative regions is systematically larger for the SLD method too, indicating a higher level of false positivity (Figure 11.9B). The ratio between the number of significant voxels in informative areas and the total number of significant voxels (the precision) is displayed in Figure 11.9C. For each of the four thresholds, the precision is higher for the FWM method.

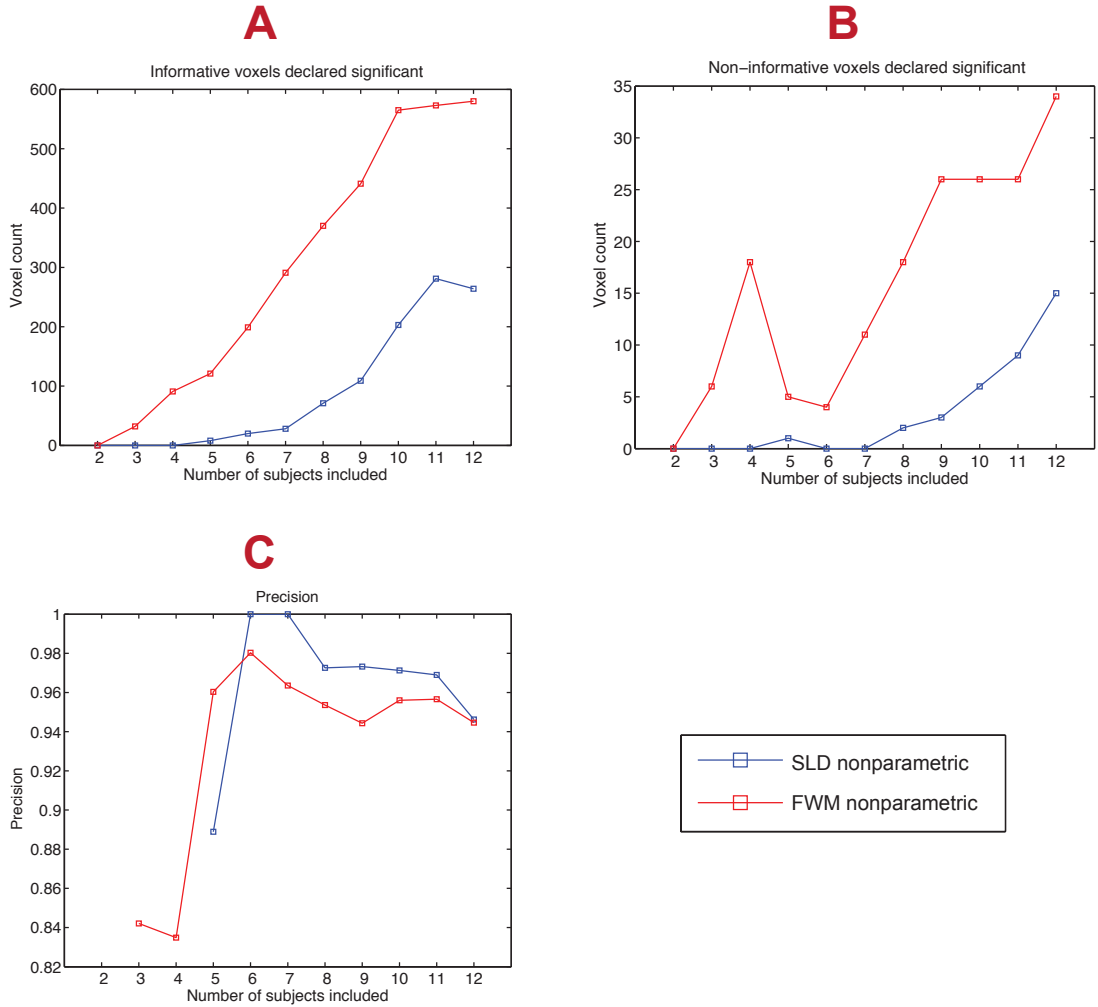


Figure 11.10: Influence of the number of subjects included for the group statistics on the SLD and FWM method in the group simulation 5cubes, shown in blue and red colors respectively. Both were computed using the the nonparametric framework. The number of virtual subjects included was varied between 2 and 12. The voxel-wise threshold for the SLD method was $p_{vox} = 0.001$, for the FWM method the threshold was set to $p_{vox} = 0.05$ (one-tailed). **(A)** For both information mapping methods, the number of true positives (number of significant voxels within the informative regions) increases monotonically with the numbers of subjects included (with the exception of the last data point in the SLD method). **(B)** The number of non-informative voxels declared significant also increases with the number of subjects included. In the case of the FWM method, however, this increase is not monotonic. **(C)** The precision of both methods is low if a small number of subjects is included and appears to minimally decrease for a higher number of included subjects.

Influence of the number of subjects To investigate the influence of the number of subjects included in the nonparametric group framework, I computed both the SLD and FWM method inclusion for only a part of the virtual subjects (between 2 and 12 subjects). For the SLD method I used a voxel-wise threshold of $p_{vox} = 0.001$, for the FWM method a (one-tailed) threshold of $p_{vox} = 0.05$. As already done in the sections before, the number of significantly labeled voxels within the informative regions was computed. This was computed for each of the 11 group sizes. The results for both SLD and FWM method are depicted in Figure 11.10. In this simulation, a higher number of included subjects generally increase the number of significant voxels in informative regions (with one exception for the SLD method when 12 subjects are included). The false positivity (non-informative voxels labeled significant), however, also increases for a larger number of included subjects. The ratio between true positives and all positives (the precision) appears to remain stationary for a medium number of subjects, but slightly decreases if all subjects are included. In the case of a small number of included subjects, the values for the precision are comparably low.

p_{cl}	0.01	0.02	0.03	0.04	0.05
expected number of clusters	1	2	3	4	5
SLD nonparametric number of clusters	0	1	1	1	1
SLD T-test number of clusters	4	7	11	13	17

Table 11.1: Number of expected clusters given type I error rate versus number of empirically found clusters for the group null simulation analyzed by the SLD method using a voxel-wise threshold of $p_{vox} = 0.001$. The first row depicts the 5 values for the type I error rate p_{cl} on cluster level that were used here. Given the total of 100 simulations, the number of expected clusters for each level of p_{cl} is displayed in the second row. In the third row, the number of empirically found clusters for the nonparametric framework is displayed. In the last row the number of empirically found clusters for the T-based method is shown.

11.2 Group null simulation

A total of 100 group null simulations were analyzed. The number of permutations was set to 100 on the single subject level for both the SLD and FWM method. The number of Monte-Carlo resampling steps was set to 10^5 in both cases. In analogy to the null simulation on single-subject level, five equidistant values for the type I error rate on the cluster level p_{cl} between 0.01 and 0.05 were used here, allowing a computation of the expectation value for the number of false positive clusters. This allowed a comparison between the number of empirically found clusters using the nonparametric and T-based frameworks and the expected number of false-positive clusters.

11.2.1 Searchlight decoding

The voxel-wise threshold for the SLD method was set to $p_{vox} = 0.001$ for both the nonparametric framework and the T-test statistics. The results of the group null simulation for searchlight decoding are shown in Table 11.1. The number of empirically found clusters for the nonparametric framework does not exceed the expected number of false positive clusters. For the T-based method, however, a systematic bias is present; the empirically found number of false positive clusters exceeds the expectation value to a high degree. This bias is large and consistent, even when allowing a certain amount of uncertainty or noise due to the low number of simulations.

11.2.2 Feature weight mapping

For the FWM method the voxel-wise threshold was set to $p_{vox} = 0.05$ (two-sided), corresponding to two one-sided tests with $p_{vox} = 0.025$. These voxel-wise thresholds were applied for the nonparametric and the T-test framework. The results of the group null simulation for feature weight mapping are depicted in Table 11.2.

Given the nonparametric framework, the number of empirically found clusters does not exceed the expected number of false positive clusters. On the other hand, the T-based

p_{cl}	0.01	0.02	0.03	0.04	0.05
expected number of clusters	1	2	3	4	5
FWM ⁺ nonparametric number of clusters	0	2	2	2	2
FWM ⁻ nonparametric number of clusters	0	1	1	1	1
FWM ⁺ T-test number of clusters	1	2670	2670	2670	2670
FWM ⁻ T-test number of clusters	0	2548	2548	2548	2548

Table 11.2: Number of expected clusters given type I error rate versus number of empirically found clusters for the group null simulation and the feature weight mapping method. The voxel-wise threshold was set to $p_{vox} = 0.05$ (two-sided). The first row depicts the 5 values for the type I error rate p_{cl} on the cluster level. Given the total of 100 simulations, the number of expected clusters for each level of p_{cl} is displayed in the second row. In the third to fourth row, the number of empirically found clusters for the nonparametric framework is displayed (for positive and negative weights). The number of empirically found clusters for the T-based method is displayed in row 5 and 6 (for positive and negative weights).

method finds a very large number of significant clusters, which drastically exceeds the expected number of false positive clusters.

11.3 3T tapping synchronization experiment

11.3.1 Searchlight decoding

The results of the searchlight decoding method analyzing the 3T tapping synchronization experiment on the group level using parametric and nonparametric statistics are shown in Figure 11.11. Both methods identify large parts of the occipital lobe (the first three slices) as significant. An involvement of these structures is highly anticipated, as visual areas are located in the occipital lobe and the classifier discriminates a sensorimotor synchronization task with flashes and a moving bar. Furthermore, secondary visual areas and the superior parietal lobule are found to be involved, which are known to be involved in visuomotor transformations and spatial attention shifts[106, 107].

The volume labeled as significant using the nonparametric statistics (Figure 11.11C) is considerably larger than for T-based statistics (Figure 11.11B). This is especially visible in the last three slices in anterior direction, where the nonparametric method decodes the motor cortex and also the lateral geniculate nucleus (LGN). Both brain structures are meaningful in regards to representing information in the visual tapping task, however are not well all identified using parametric statistics. The empirical voxel chance distributions differ greatly across locations. Figure 11.11D shows the accuracy for which the area of the (normalized) chance distribution is $<0.1\%$, therefore depicting the accuracy level for $p_{vox} = 0.001$ (this map served as threshold map for the cluster search).

In total, 16794 voxels were labeled as significant using the proposed nonparametric method, while only 9257 voxels were found significant for the T-based approach. 8894 voxels were labeled as significant by both methods, leaving 363 voxels identified when solely using the T-test framework and 7900 voxels by only the proposed nonparametric method.

The empirical cluster-size histogram is displayed in Figure Figure 11.12. The minimum cluster size was 7 voxels in this data set for the given voxel-wise threshold $p_{vox} = 0.001$ and a cluster p -value of $p_{cl} = 0.05$.

Testing on a voxel-wise level for normality of the decoding accuracies using the Shapiro-Wilk test (see Section 5.2.4), a total of 7725 voxels is did not follow a normal distribution (on a significance level $\alpha = 0.05$). In other words, the accuracy values of at least 13% of all locations is not distributed normally (the total number of voxels was 59329).

11.3.2 Feature weight mapping

The raw feature weights of the fMRI tapping study are displayed in Figure 11.13A. Note that the weights can take both positive and negative values. Figure 11.13B depicts the parametric T-test method including Gaussian random fields and cluster level FDR. The Figure shows a double overlay of a right-tailed T-test (for positive weights) and the left-tailed T-test (for negative weights). The results of the proposed nonparametric method are shown in Figure 11.13C.

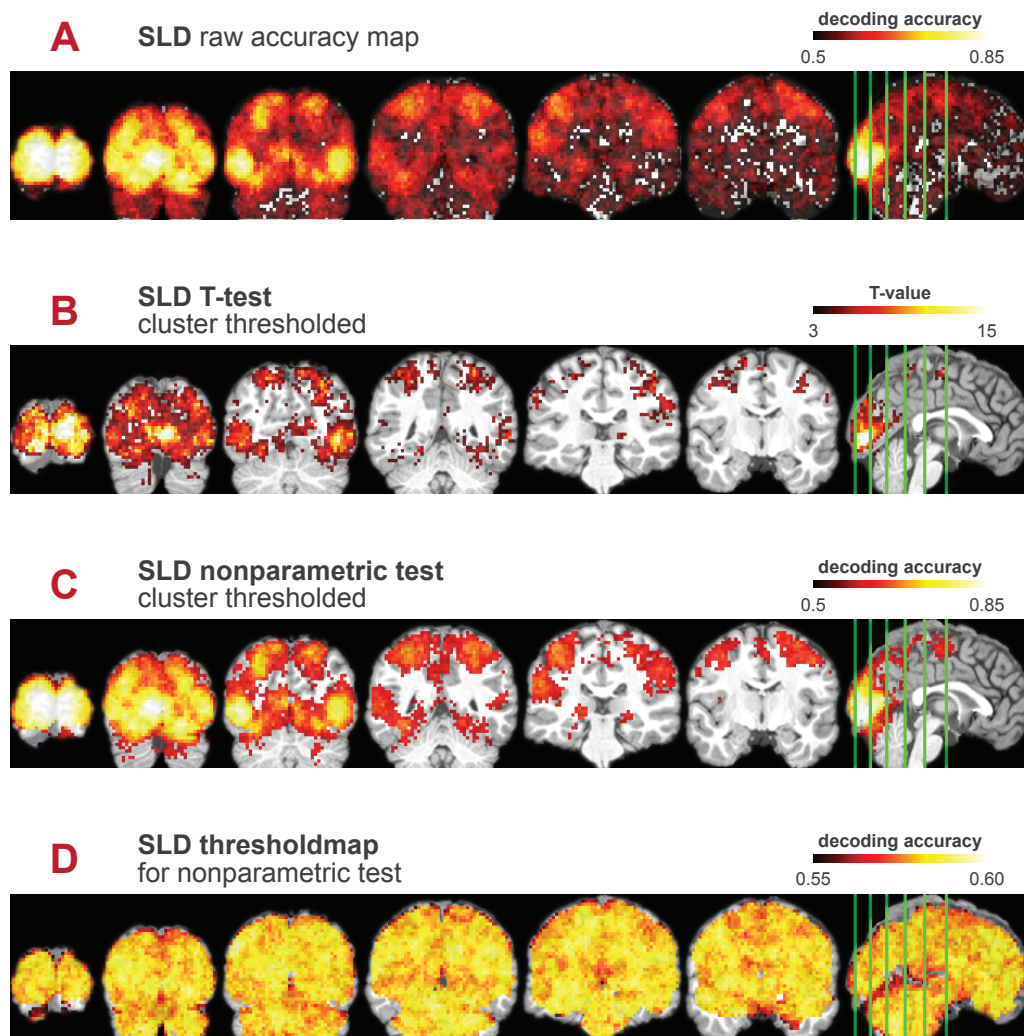


Figure 11.11: Comparison of the group-level searchlight decoding results of the 3T tapping synchronization experiment between the nonparametric method and standard T-based methods. **(A)** Raw decoding accuracy map without further multiple comparisons corrections. **(B)** The results of a standard T-test, corrected for multiple comparisons using Gaussian random field and FDR methods on cluster level as implemented in SPM8, with a voxel-wise threshold of $p_{vox} = 0.001$. **(C)** Results of the new nonparametric method implementing the multiple comparisons correction on cluster size level, using the same $p_{vox} = 0.001$. **(D)** Threshold map for the nonparametric cluster search algorithm, revealing the inhomogeneity of the local chance distribution.

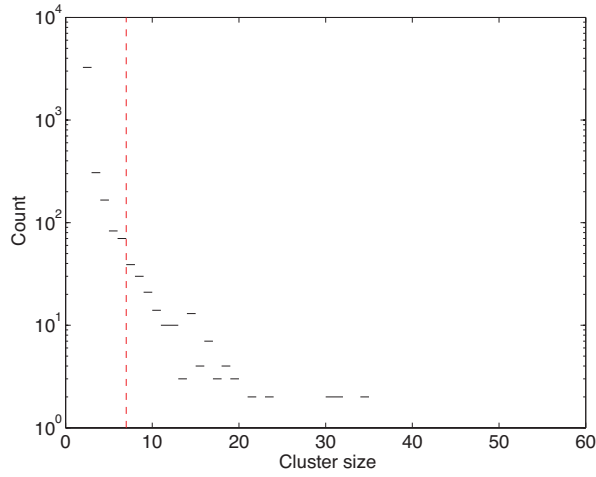


Figure 11.12: Cluster-size histogram of the group-level analysis of the 3T tapping synchronization experiment for the nonparametric analysis using searchlight decoding and a voxel-wise threshold of $p_{vox} = 0.001$. The red line marks the $p_{cl} = 0.05$ percentile (uncorrected) of the cluster size distribution (cluster-size = 7 voxels).

In terms of positive weights, both the T-based and nonparametric method appear comparable. For the negative weights, however, the T-based method labels areas within white matter as significant (see Figure 11.13B, slice number 4), indicating false-positivity. The nonparametric method, on the other hand, identifies a early-visual region (Figure 11.13C, first slice) that is not found significant with the T-based method. The inhomogeneity of the empirical null distribution for positive and negative weights are displayed in Figure 11.13D and E respectively.

In total, 5063 voxels are labeled as significant using the T-based method. The nonparametric method yields 3803 significant voxels. 2872 voxels were determined as significant by both methods, leaving 2191 voxels identified when using solely the T-test framework and 931 voxels by only the proposed nonparametric method .

The two empirical cluster-size histograms are displayed in Figure 11.14. The minimum (uncorrected) cluster size was 4 voxels for both positive and negative weights in this data set for the given voxel-wise threshold of $p_{vox} = 0.05$ and a cluster p -value of $p_{cl} = 0.05$.

The voxel-wise test for normality of the weights (Shapiro-Wilk test, see Section 5.2.4) found that 8587 voxels did not follow a normal distribution (on a significance level $\alpha = 0.05$). Therefore, the weights of at least 14% of all locations are not distributed normally (the total number of voxels was 59329).

11.3.3 Comparison of SLD vs FWM

The results of the fMRI experiment for the searchlight decoding method (SLD) and the feature weight mapping (FWM) approach are displayed for a direct comparison in Figure 11.15. Both methods were corrected for multiple comparisons using the proposed nonparametric statistics each using two different voxel-wise thresholds of $p_1 = 0.05$ and $p_2 = 0.001$.

Similar as in the group simulations before (see Section 11.1.2 on page 107), the SLD

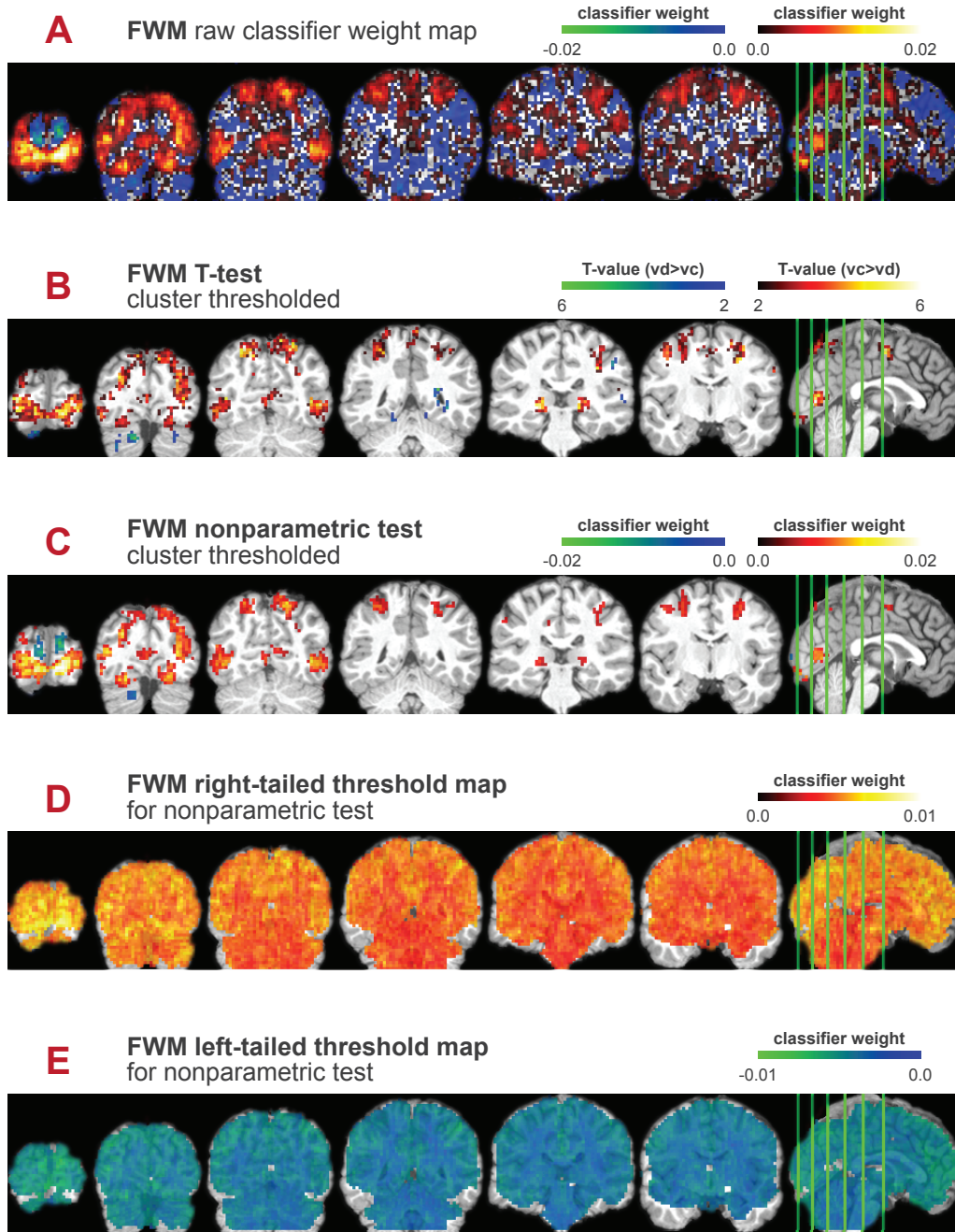


Figure 11.13: 3T tapping synchronization experiment analyzed by the feature weight mapping method using parametric (T-based) and nonparametric frameworks. **(A)** Mean feature weights over all 12 subjects without any further correction. **(B)** Results of a (two-tailed) T-test applied on the feature weights, corrected for multiple comparisons by usage of Gaussian random fields and FDR on the cluster p -values. The initial voxel-wise threshold was set to $p_{vox} = 0.05$ for the two-tailed test, which is equivalent to two one-tailed T-tests at a threshold level of $p_{vox} = 0.025$. $vd > vc$ (in the color bar) indicates the left-tailed T-test here (where the weights of condition $vd = \text{visual discrete}$ are higher than $vc = \text{visual continuous}$) and $vc > vd$ vice versa (right-tailed test). **(C)** Proposed nonparametric statistics based on random permutations and Monte-Carlo resampling implementing a multiple comparisons correction using the same initial voxel-wise threshold of $p_{vox} = 0.05$. Also here, effectively two one-tailed tests with a voxel-wise $p_{vox} = 0.025$ had been carried out. **(D)** Positive threshold map for the nonparametric method, weights that surpass the voxel-wise threshold are labeled as supra-threshold voxels for the cluster search. **(E)** Negative threshold map, weights that are below the voxel-wise value are counted as supra-threshold voxels.

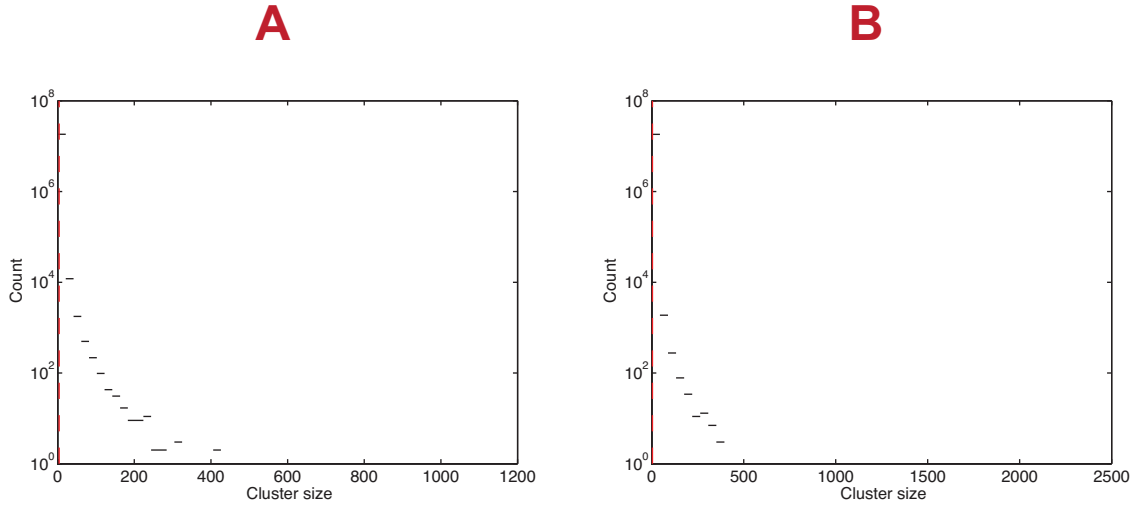


Figure 11.14: Cluster-size histogram of the 3T tapping synchronization experiment for nonparametric analysis based on feature weight mapping and a voxel-wise threshold of $p_{vox} = 0.05$. The red line marks the $p_{cl} = 0.05$ percentile (uncorrected) of the cluster-size distributions. **(A)** Cluster-size histogram for clusters with *positive* weight (minimum cluster size = 4 voxels for the $p_{cl} = 0.05$ percentile). **(B)** Cluster-size histogram for clusters with *negative* weight (minimum cluster size = 4 voxels for the $p_{cl} = 0.05$ percentile).

method labels a considerably higher volume as significant (given a fixed voxel-wise threshold p_{vox}) while the results of the FWM method, are more fine-grained. However, both the visual areas and other cortical (motor cortex) and subcortical structures (such as the lateral geniculate bodies in 11.15 at slice 5) are much more clearly delineated using the FWM method and appear inflated for the SLD method.

In contrast to the SLD method, the FWM method is able to discriminate the sign of the classification decision in the visual cortex (see Figure 11.15, first slice): While the FWM method finds regions of the primary visual cortex (Brodmann area 17) with positive weights, other regions in visual association areas (Brodmann area 18) are labeled with negative weights. It is noteworthy, that the SLD method finds the highest decoding accuracies in the border region in between.

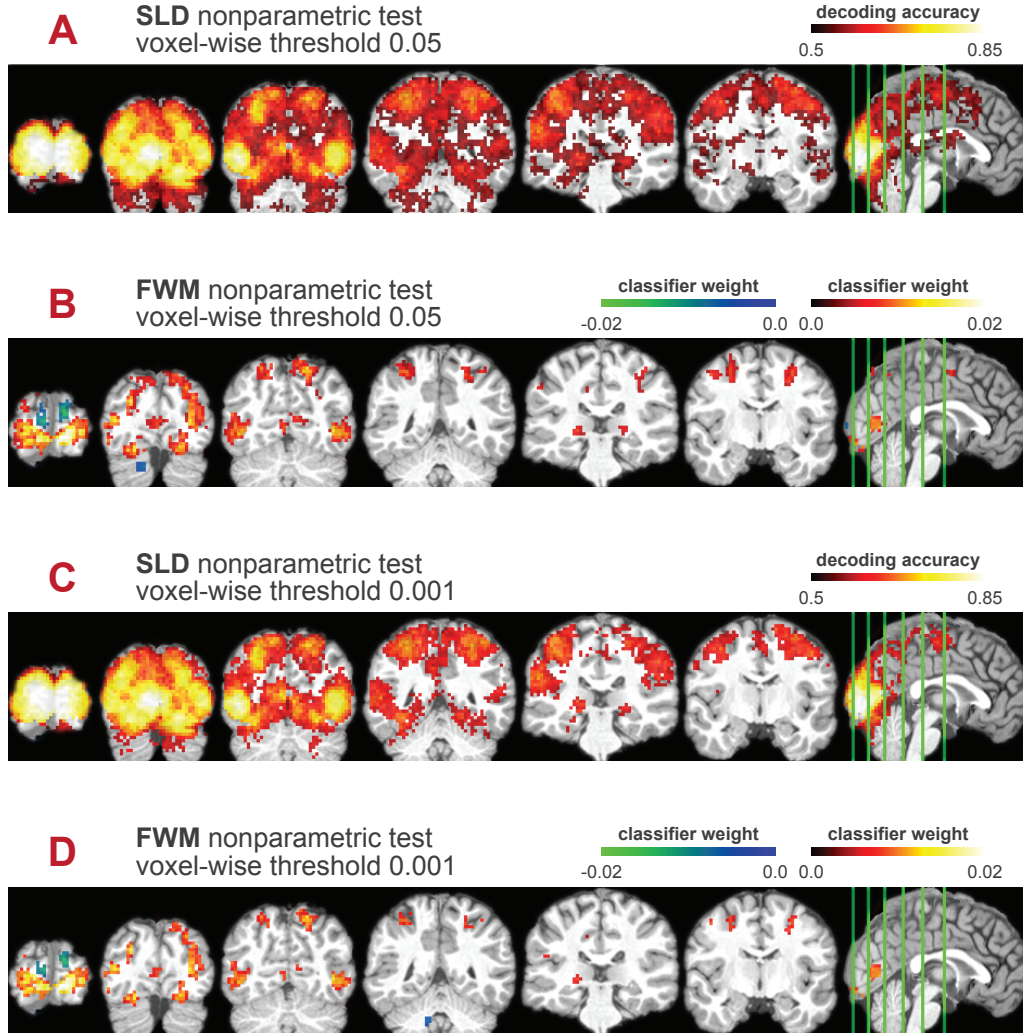


Figure 11.15: Comparison of the FWM and SLD results of the 3T tapping synchronization experiment, both using the nonparametric framework for the multiple comparisons correction. The initial voxel-wise threshold had been varied between two values ($p_{vox} = 0.05$ and $p_{vox} = 0.001$). (A) Searchlight decoding map, analyzed with an initial voxel-wise threshold of $p_{vox} = 0.05$. (B) Feature weight mapping method using the same initial voxel-wise threshold of $p_{vox} = 0.05$. (C) Searchlight decoding map implementing the higher voxel-wise threshold of $p_{vox} = 0.001$. (D) Corresponding feature weight map with the higher threshold $p_{vox} = 0.001$.

Chapter 12

General results

12.1 Cross-validation influence simulation

The aim of the cross-validation influence simulation was to show that the theoretical estimation of the classifier's significance level using a binomial model can only be applied if independency is given, and that the empirical *deviation* to the approximation indeed depends on the degree of correlation between the binomial variables. The results of the simulation are depicted in Figure 12.1A, where I computed the histogram of decoding accuracies for each of the six scenarios (no cross-validation, 2-cv, 5-cv, 10-cv, 20-cv, 50-cv). Additionally, I plotted the binomial distribution with parameters (200,0.5) using black dots. Figure 12.1B displays a magnification of the grey dotted rectangular area of Figure 12.1A. The binomial model is exact only if no cross-validations are applied. The empirical distributions using cross-validation have a substantially higher variance compared to the binomial. Furthermore, the variance of the null distribution is clearly dependent on the number of applied cross-validations; the more cross-validations that are applied, the higher the variance of the null distribution. The deviation between binomial and empirical distributions (using the error term described in Equation 7.1 on page 58) is shown in Figure 12.1C. If no cross-validation is applied, the difference in the distributions becomes zero, i.e. the binomial model is indeed exact. If cross-validation is applied, the deviation monotonically increases for a higher number of applied cross-validations.

12.2 Simulation undersampling the permutation space

The goal of the simulation undersampling the permutation space was to investigate the impact of undersampling on the group null distribution of decoding accuracies. For this, a varying number of permutations on the single-subject level were carried out. In total, four numbers of single subject permutations were computed: 10, 100, 1000 and 10000. Each undersampling step was repeated for 1000 times, each time a group null distribution was computed using the nonparametric framework (permutations and Monte-Carlo resampling), followed by a normal fit and estimation of the two parameters μ and σ .

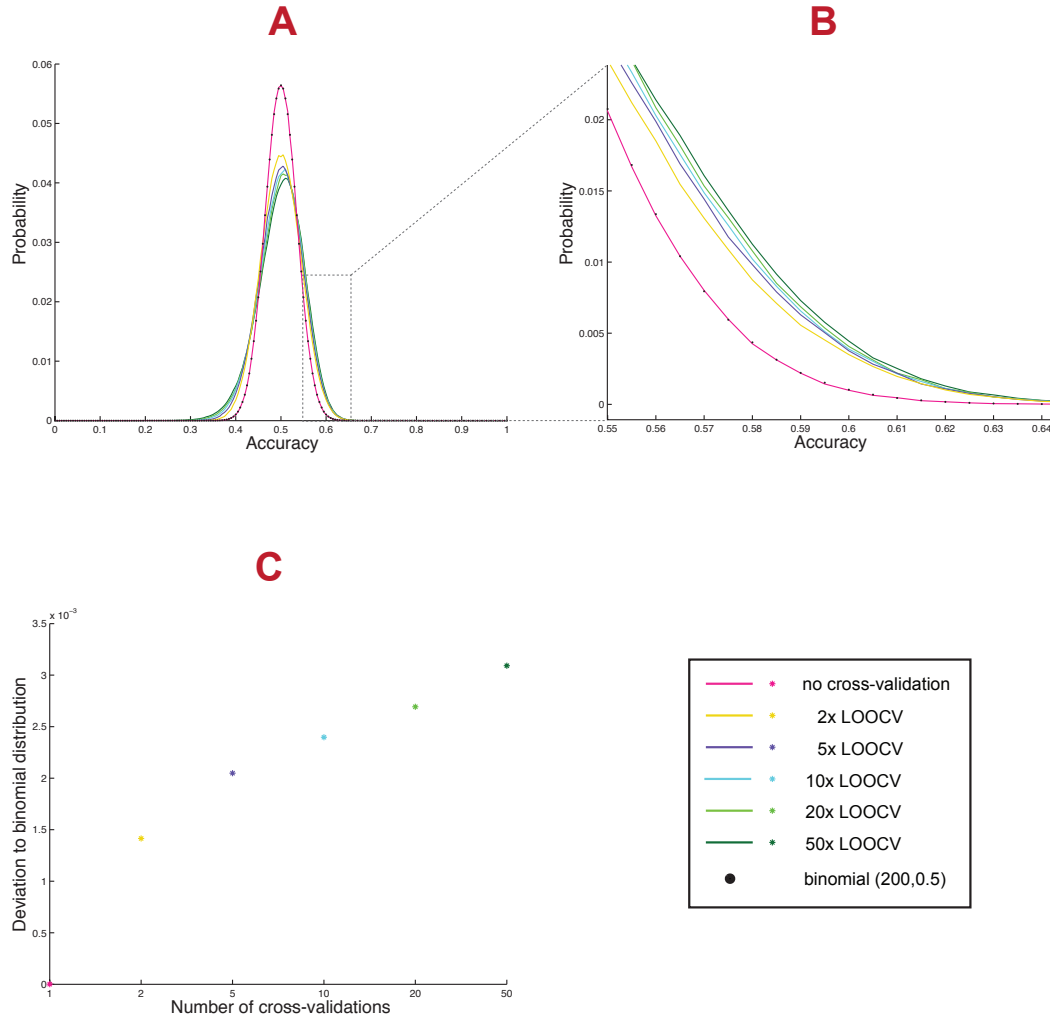


Figure 12.1: Cross-validation influence simulation showing the impact of different cross-validation schemes on null data sets. **(A)** Six scenarios with different data partitions and hence number of cross-validations are displayed here, ranging from no cross-validation to 2-, 5-, 10-, 20-, or 50-fold cross-validation. The theoretical distribution assuming independency is depicted by black dots. **(B)** Zoomed-in image of the dotted grey area of the above figure. It is clearly visible that the distributions get wider if more cross-validations are applied. **(C)** Deviation of the empirical distributions to the binomial distribution (200, 0.5). If no cross-validations are applied, the distributions are matched exactly. The more cross-validations that are used, the monotonically higher the deviation to the binomial becomes.

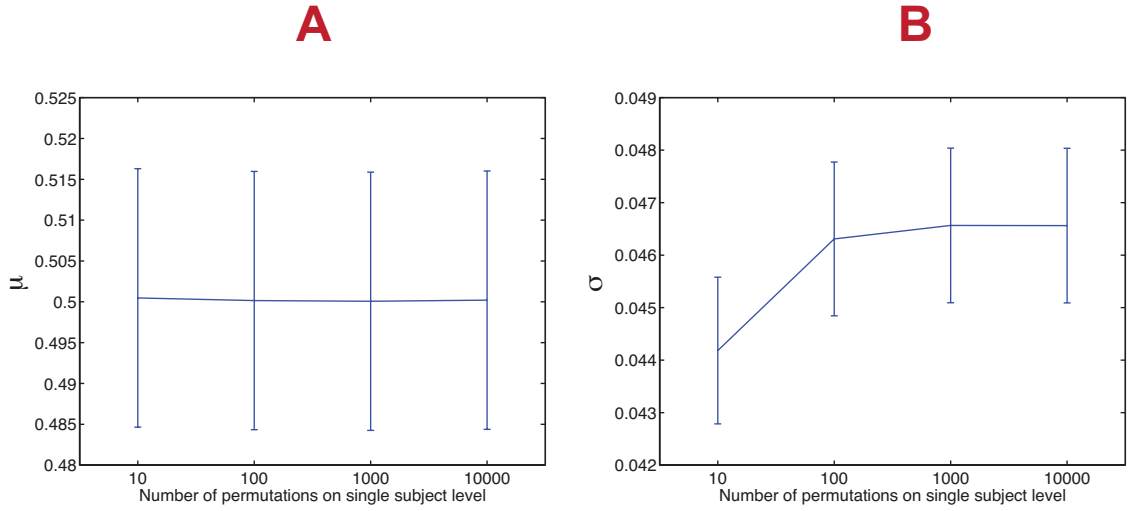


Figure 12.2: Results of the simulation undersampling the permutation space on the group level distributions of accuracies. In total, four levels of available permutations on single-subject level were present. **(A)** Estimated mean value μ of the group level distribution of accuracies. The error bars depict the standard error of the 1000 repetitions. In regards to the mean value μ , the number of available permutations on the single-subject level does not appear to have influence **(B)** Estimated parameter σ of the group level distribution of accuracies. Also here the error bars are the standard error of the 1000 repetitions. The estimation of the parameter appears convergent for 100 available permutations carried out on the single-subject level.

The results are shown in Figure 12.2. As is visible in Figure 12.2A, there is no considerable difference in the estimated mean parameter μ of the normal distributions depending on the number of available permutations on single subject level. The standard deviation parameter σ , however, displayed a dependency on the number of permutations on single subject level. Given 100 or more permutations, the parameter σ converges (see Figure 12.2B). The error bars in Figure 12.2 represent the standard error of the 1000 repetitions of the simulations for μ and σ respectively, and was calculated as $SD_{\mu} = \frac{\hat{\sigma}_{\mu}}{\sqrt{N}}$ or, respectively $SD_{\sigma} = \frac{\hat{\sigma}_{\sigma}}{\sqrt{N}}$, where $\hat{\sigma}$ is the standard deviation of the 1000 repetitions and $N = 1000$.

The results indicate that on the single-subject level, at least 100 permutations should be computed to ensure stability of the group null distribution. Furthermore the results show that 100 permutations are a sufficient number for constructing the group distribution.

Part IV

Discussion

Chapter 13

Statistics in fMRI

In my thesis I have introduced a framework for computing nonparametric statistical inference for classification-based fMRI. I have shown that the nonparametric framework is applicable to two different information mapping methods, namely searchlight decoding and feature weight mapping. In particular, I have extensively compared the nonparametric framework to parametric alternatives using simulations and actual fMRI data sets. In this section, I want to outline the theoretical reasoning for why nonparametric statistics are a better choice for dealing with classification-based data as compared to parametric alternatives. Furthermore, I want to discuss the features and characteristics of the proposed nonparametric framework.

13.1 Pitfalls of parametric statistics in classification-based fMRI

For deriving statistical inference for classification-based fMRI on the group level, parametric statistics are often employed. The two most widely-used parametric frameworks here are T-based methods (which can be applied to accuracy and weight maps)[102, 104, 105] and theoretical approaches based on Binomial models (which can be applied to decoding accuracies)[108]. In the following, I want to discuss the pitfalls of these two representatives of parametric methods in the context of classification-based fMRI.

13.1.1 T-based statistics

The Student's T-test is the most commonly practiced method for determining the probability of a decoding result on the group level[5]. Furthermore, the same method can also be applied on feature weights for the group-level inference. Importantly, T-tests impose certain assumptions on the data. For instance, the samples need to be distributed normally, particularly if the sample size is small. Furthermore, the underlying distribution from which samples are drawn should be continuous. Both requirements are problematic if a T-test is performed for a second-level group analysis of decoding accuracies. In general, decoding accuracies are not normally distributed, because the unknown distribution of decoding accuracies is generally skewed and

long-tailed. In practice, the distribution depends heavily on the classifier that is used and the input data itself. I have shown that in practice a significant part of the experimental fMRI data, analyzed with the SLD or FWM approach, exhibits non-normality (at least 13% or 14% respectively, see Section 11.3 on page 115).

Moreover, in the case of searchlight accuracies, the samples are not drawn from a continuous distribution. The indicator function, which maps the number of correctly predicted labels to an accuracy value between $[0, 1]$, can only take certain values: for k cross-validation steps and a test set of size t_{ts} , only $k \cdot t_{ts} + 1$ different values between $[0, 1]$ can be taken. Furthermore, the high variance on the single-subject accuracies depicts a problem for T-based frameworks. One of the most critical assumptions of any classification-based method is that the observations are drawn independently from the data set[46]. This imposes a problem for the typical fMRI data set, given the severe temporal contamination of subsequent scans due to autocorrelation[109]. Therefore, the prerequisite of independence has to be approximated by taking into account well-separated groups of scans or their statistical estimation parameters (e.g., from a general linear model on separate trials, as done in this work). Hence, the ultimate number of observations available for classification is greatly limited. This limitation of samples imposes severe challenges, as demonstrated in recent work[110]. Most noteworthy, an inverse relation between the number of samples and the variance of accuracies is found; the fewer samples used, the larger the variance in the estimated accuracies. The variance is also driven by the size of the test set, as had been shown previously[111]: The smaller the size of the test set, the greater the variance in the estimated accuracy. It is important to emphasize that the variance of the estimation of accuracies must not be confused with the true underlying variance of the performance of the classifier (caused, for example, by inter-session and inter-subject variability and generally non-observable in a real data set). However, simulations that enable a measurement of true performance clearly demonstrate that the variance is, in fact, dominated by the effects of a small sample size [110]. Furthermore, the indicator function mapping the number of correctly predicted labels to an accuracy value between $[0, 1]$ is of a discrete nature, which additionally increases the variance for small data sets[112]. For these reasons, a statistical inference method for classification-based decoding in fMRI should not heavily rely on the variance of the decoding accuracies on a single-subject level. However, the commonly practiced T-tests on single subject accuracies for carrying out group inference fundamentally implement the variance, as the square root of the variance enters the denominator of the T-formula (see equation 5.5 on page 35)[5].

13.1.2 Binomial models

For determining the statistical significance of a *decoding accuracy* on the single-subject or group level, binomial models can be employed. For the sake of simplicity, I will base my argument first on single-subject analysis and generalize it later on for group-level statistics. This implies a binomial draw with the parameters (N, p) , where N is the number of samples whose labels are predicted (i.e. the size of the test set) and p is set to $1/2$ in a two-class paradigm (see Equation 5.7 on page 36). If cross-validation schemes are applied, however, the situation becomes more complex. Given k leave-one-out cross-validation folds, it is often

the practice to treat the k cross-validations as one single classifier[47]. The rationale behind this is that each cross-validation fold yields one binomial random variable, hence the whole cross-validation process yields k identically distributed binomial random variables X_1, \dots, X_k each with the parameters (N, p) . As the accuracy is defined as the total number of correctly classified samples over all the cross-validation folds (divided by a constant), the *sum* of these k binomial variables is computed. Under the assumption of independency, this sum again is a binomial random variable, however with the parameters $(N \cdot k, p)$ [113, page 214]:

$$f_Z(c) = \binom{N \cdot k}{c} p^c (1 - p)^{N \cdot k - c} \quad (13.1)$$

In other words, the k cross-validations are treated as one single classifier, which estimates $N \cdot k$ labels—under the assumption of independence of the respective binomial random variables.

The independence of *test* examples in each cross-validation fold, however, does not automatically assure independence of the binomial random variables X_1, \dots, X_k . Most importantly, the correlations between training sets and testing sets in different folds cause the binomial random variables to be correlated to each other. This correlation violates the earlier assumption of independence, and renders the above used procedure for summing independent binomial variables formally incorrect[5].

The empirical results revealed that the deviation from the theoretically derived sum Z monotonically depends on the degree of correlation between the binomial random variables X_1, \dots, X_k ; the higher the correlation between cross-validation folds, the larger the deviation to the single classifier approximation (see Figure 12.1 on page 122). On the contrary, I showed that in the case of no cross-validation and a true single classifier estimating of $N \cdot k$ labels, the binomial model fits exactly. Moreover, the variance of the empirical null distributions depends on the degree of correlation between cross-validation folds; the higher the correlation, the broader the distribution. When applied in statistical inference, the smaller variance of the null distribution from the binomial model has an effect of *overestimating* the p -values. Effectively, smaller p -values will be reported from the binomial model, than from the empirical ones, as reflected in studies[47]. Therefore, adopting the above binomial model in the case of correlated cross-validation principally increases the false positivity[5].

The argument can be generalized to the group level, as the accuracy of one single voxel on the group level could be modeled by a single classifier estimating $N_{sub} \cdot N \cdot k$ labels, where N_{sub} is the number of subjects. However, in case cross-validation procedures are applied on a single subject level, the same argument holds here too: it is formally incorrect to model multiple cross-validation steps as single classifier, regardless of the independence between subjects.

As a side note, it should be mentioned that the issue of correlation between cross-validation steps may be mitigated by estimating the correlational structure between the cross-validation folds and application of *correlated* binomial models[114]. The improved model, however, then relies on additional assumptions, e.g. about the reliability of estimation of the correlation between cross-validation steps. Furthermore, the binomial model does not provide a

solution to the multiple comparisons problem; it merely provides a map of uncorrected p -values. As the number of voxels is very large, a correction is a necessity, requiring the development of a further framework (e.g. local FDR methods[115, 116] or an adaptation of Gaussian random field methods).

13.2 Characteristics of the nonparametric framework for classification-based fMRI

Nonparametric tests, such as the permutation test, are exact regardless of whether the underlying distribution is normal or not. Furthermore, permutation tests rely on minimal assumptions[74, page 23], such as exchangeability of the data points. It should be noted that this assumption is imposed not only for permutation tests but also for classification methods in general[46]. The theoretical applicability of permutation tests for classification-based methods is well established[117]. A limitation for the usage in fMRI studies is that the available number of permutations on the single subject level is relatively small, since the number of independent observations is limited. Effectively this enforces procedures which maximize the number of data points while ensuring independency, such as estimating one GLM-derived β -estimate per experimental trial. Since the underlying numbers of data points (time steps, i.e. scans) which are used for obtaining the β -estimates are rather small (in the case of the fMRI experiments used in my study 9 or 8 subsequent scans were used for the 3T / 7T study respectively), it is likely that the β -estimates are subject to a large variability. Ultimately, the classification results of single subject studies may well be increased if the number of experimental trials was larger, as this would allow reasonable permutation based statistics using β -estimates derived from multiple trials (in contrast to one β -estimate per experimental trial); this would effectively decrease that variance of the data points and may improve the classifier’s model.

On the group level, this limitation is less severe; the Monte-Carlo resampling technique circumvents the issue of a low number of available permutations on the single subject level. Furthermore, the number of permutations necessary for each subject is comparably small; 100 permutations are already sufficient to construct a large pool consisting of up to 10^5 resampled group chance maps (see Figure 12.2 on page 123).

Finally it should be noted that the proposed permutation scheme avoids potential biases due to an uneven distribution of samples across classes (i.e. if one class dominates the training or test subset). In the case of the searchlight decoding method, the correlational structure between all cross-validations is also preserved, as the data is permuted before the cross-validation scheme is applied and the permutation is held fixed for all cross-validation folds.

13.2.1 Preservation of spatial structure in chance maps

The preservation of spatial structure is fundamental to the nonparametric statistical framework, since it incorporates the analysis of the spatial structure in the *chance maps* (either on the

single-subject level or group level). Most critically, this allows one to assign probabilities to these features. The same analysis of spatial features is then carried out in the non-permuted (original) maps. By recourse of the probabilities which were previously derived from the chance maps, probabilities can ultimately be assigned to the spatial features in the original maps. If the spatial correlations of the chance maps were not preserved, the obtained probabilities for spatial features would either over- or underestimate the ground-truth probabilities of the original maps. Effectively this would introduce a bias, which would imply high levels of false-positivity or false-negativity. The proposed nonparametric framework holds one permutation of the order of data points fixed for all voxels per chance map. This ensures that all spatial correlations which had been present in the underlying data points are preserved in the resulting chance maps.

Furthermore, in the case of the searchlight decoding implementation, the method itself acts as a spatial filter. Hence spatial correlations are introduced by the SLD method itself, which depend on the diameter (and geometry) of the searchlight. By virtue of keeping the permutation fixed for all locations, this source of spatial correlations is also fully reflected in the chance accuracy maps.

The spatial structure is preserved not only on the single-subject level (i.e. the permuted chance maps) but also on the group level (i.e. in the Monte-Carlo resampled maps). On the group level, however, the combined correlational structure is preserved in regards to the spatial dimensions on the *group level*. The degree of spatial dependency may be smaller as compared to the single subject level due to inter-subject variations.

13.2.2 Threshold map procedures

The proposed nonparametric approach incorporates the spatial inhomogeneity of the null distributions in the cluster-search algorithm. If the algorithm used a constant threshold level (e.g. one global accuracy level for all locations in the SLD method or one size for the weights in the FWM method), the resulting local cluster sizes would be biased: in the case of a local null distribution which is more broad, the local cluster size would be overestimated, while in the case of a local chance distribution which is more narrow, the cluster sizes would be underestimated. Hence, the overall cluster size distribution would be biased, depending on the degree of spatial inhomogeneity and the choice of the global threshold.

The threshold map procedure on a single-subject level is effectively equivalent to performing a permutation test on every voxel and use the results to apply a threshold. This is the case because only voxels exceeding a certain statistical threshold defined by the permutation distribution do surpass. In the case of group-level studies, the threshold map procedure is equivalent to a combined permutation and Monte-Carlo resampling test, where a statistical voxel-wise threshold is derived.

Naturally, the width of the permutation or Monte-Carlo resampled distributions depends on several factors such as the actual information content; in the presence of information, the width and possibly location of the nonparametric distribution is increased or shifted to the right. Hence the threshold at this location is rendered more *conservative*. Nevertheless, in the

presence of information, the nonparametric procedure does not become overly conservative, as on the group level it is possible to compare the obtained sensitivities to parametric methods; the sensitivity of the nonparametric framework is superior here (for a further discussion and comparison see Section 13.3 on page 134). The same argument holds for the opposite case where no information is present. Here, the permutation/Monte-Carlo resampling distribution does not become exceedingly permissive to false-positives, as demonstrated in the null simulations for validation on both single-subject and group level (see Section 10.2 on page 87 and Section 11.2 on page 113).

13.2.3 Cluster statistics

On the voxel-level, the multiple comparisons problem is rather severe, considering that up to 50000 tests are carried out for 3 Tesla data and 500000 tests or more for 7 Tesla data. This issue is mitigated by performing the statistics on the spatial features of the underlying images. A simplistic definition of spatial features describes them as spatially connected areas surpassing a voxel-wise threshold, in other words clusters. Importantly, if cluster statistics are used, the fundamental units of interest are regions and not voxels[118]. The principal advantage is that if clusters are considered instead of voxels, the statistical tests are carried out on the clusters themselves (e.g. by assigning probabilities to cluster sizes). This reduces the number of tests carried out by several orders of magnitude. The absolute number of tests on the clusters, however, highly depends on the voxel-wise threshold that is used for determining whether a connected voxel can be joined into a cluster or not. Furthermore, it is worth mentioning that cluster-based approaches have been demonstrated to be statistically more powerful than voxel-based tests[84].

The basic idea of cluster size inference is to exploit the fact that the probability of two voxels exceeding a given voxel threshold and simultaneously being contiguous is *smaller* than the chance of one sole voxel surpassing a threshold[119]. The determination of the *probability* of a cluster of a given size, however, is not trivial and also depends on the degree of spatial correlations. For the group level¹ this can be achieved by using (T-based) random field methods, which enable a mapping between cluster size and probability (see Section 5.4.3 on page 44). Most importantly, the random fields procedure critically depends on the correct estimation of the underlying smoothness of the accuracy or weight maps, as the probability for a given cluster size is a function of the estimated smoothness (see Equation 5.8 on page 45). It is questionable, however, that the algorithms for the estimation of smoothness can be directly applied to classification results (accuracy or weight maps), in particular for two reasons: firstly, the application of the underlying T-based statistics is formally incorrect here. Secondly, no prior spatial smoothing is carried out for the pattern based analysis, which is problematic since random field method have been shown to perform poorly in case of not sufficiently smooth images[84].

The nonparametric framework introduced in this work does not rely on an explicit estimation of image smoothness for the derivation of a mapping between cluster size and prob-

¹Principally, T-based methods may also be applied on the single subject level if cross-validation is carried out. The discussed problems regarding T-based methods, however, become more pronounced here.

ability. The image smoothness is considered *implicitly* in the formation of the empirical cluster size records of the chance maps. This is possible, as spatial correlations are preserved in the chance maps. Hence, the size of underlying spatial correlations and image smoothness is directly reflected in the empirical cluster-size histograms (see Figure 10.6 on page 88). As an direct estimation of smoothness is not necessary, it can be stated that the nonparametric cluster size thresholding is the more elegant and versatile solution, however at the cost of larger computation times.

Lastly, it should be noted that a correction for the multiple comparisons problem has to be carried out on the derived cluster p -values, regardless whether these values had been obtained by random field methods or computed by nonparametric means. As stated before, this multiple comparisons problem is comparably mild, as the number of clusters in the original maps usually is in the range of hundreds. Therefore, standard FDR methods are applied here in both the T-based and the nonparametric framework.

13.2.4 Dependency on the voxel-wise threshold

The results of both the parametric and nonparametric cluster-based approaches depend highly on the choice of the initial voxel-threshold, which effectively determines whether a voxel is joined into a cluster or not. Since the threshold is a free parameter which cannot be deductively derived from first principles, the choice of this threshold principally underlies a certain arbitrariness of the experimenter. On one side of the spectrum, thresholds that are too high (i.e. low p -values) drastically reduce the sensitivity (while increasing the precision); while on the other side thresholds that are too low (i.e. high p -values) hurt the localization of the truly informative regions as the sensitivity becomes high but the precision low². In terms of a worst-case scenario, lowering the initial voxel threshold might even lead to the merging of otherwise separated clusters, which in turn is reflected in the empirical null distribution of clusters. By means of this mechanism, clusters in the original maps may retrieve an insignificant probability. Indeed this may have been the case in for the two fMRI studies on a single subject using searchlight decoding (see Section 10.3 on page 88 and Section 10.4 on page 93), where no results were labeled as significant for the lowest threshold of $p_{vox} = 0.05$ (and hence a threshold of $p_{vox} = 0.01$ had to be used instead).

However, based on the simulations (see Figure 10.5 on page 85 and 11.9 on page 110), a reasonable choice of voxel-wise thresholds which ensures a good degree of comparability could be found heuristically: for the SLD method, this threshold is determined as $p_{vox} = 0.001$, the corresponding voxel-wise threshold for the FWM method is then $p_{vox} = 0.05$. Both values were selected on the basis of a comparable sensitivity and precision. However, the sensitivity and precision of both methods highly depends on the geometric distribution and intensity of the signal (see Figure (10.5)). Given this, the choice of a voxel-level threshold can only be an approximated in a heuristic fashion.

An alternative method is to eliminate the choice of the voxel-wise threshold by substi-

²This is a principal trade-off problem inherent to frequentist statistical methods, I have illustrated the problem already in the introduction in Figure 5.1 on page 33

tution with other parameters that themselves are optimized by receiver-operating-characteristics³ for a certain range of signal and noise characteristics[120]. While this method appears to be more stable in regards to the choice of the (optimized) parameters, it is questionable whether the range of signal characteristics can appropriately reflect searchlight decoding maps, since here the signal characteristics depend highly on the searchlight parameters (such as the diameter).

13.3 Comparison between nonparametric and parametric statistics in classification-based fMRI

In the following I want to put aside the theoretical concerns on parametric T-based statistics for classification-based fMRI and compare the parametric with the nonparametric approach from a pragmatic and practical point of view. The comparison is oriented on the quality of each statistical testing method, which I measure in a three-fold manner: Firstly, the overall *sensitivity* of the method is of interest (which was defined as the fraction of true positives in *all* positives, see Section 5.1 on page 31). The measurement of the sensitivity can be carried out in a quantitative way using simulations, where the actual ground truth (i.e. informative regions) is known. Additionally, in a more indirect and qualitative sense, the sensitivity of a test can be estimated by real fMRI data: if it is highly plausible from a neuroscientific point of view for a certain brain region to show involvement in a specific task and these brain regions are identified as informative by one method but not another, then the sensitivity of the former method can be presumed as higher.

The second measure of interest for determining the quality of a statistical test is its *precision* (which was defined as the fraction of positively labeled tests that were in fact ground truth positives, see Section 5.1). Similarly as before, a quantitative measure of precision can be assessed by virtue of simulations, where the ground truth is known. Furthermore it is possible to get an impression of the precision when using fMRI data: if it is highly *implausible* from a neuroscientific point that certain brain region exhibits involvement in a task (e.g. white matter, cerebrospinal fluid) and one statistical analysis method labels systematically more voxels as significant in these regions, it can be concluded that the precision of this method is likely *lower*.

The third criterion for measuring the quality of the statistical test is the credibility in terms of overall type I error control. Using a large number of null simulations, where *every* significant result represents a false positive, the expected error rate can be cross-checked with the empirical error rate.

13.3.1 Sensitivity

For the group simulation, the sensitivity could be defined as the fraction of informative voxels (ground truth positives) which were correctly labeled significant. When comparing the

³receiver-operating-characteristics are plots between the true positive rate against the false positive rate.

sensitivity of the nonparametric and the parametric method for searchlight decoding, the nonparametric method outperforms the T-based approach on all voxel-wise thresholds. When using the optimal voxel-wise threshold $p_{vox} = 0.001$ for the SLD method, the gain in sensitivity is about 100% for the nonparametric framework as compared to the T-based method (see Figure 11.3 on page 102). The difference in sensitivity between nonparametric and parametric tests for the feature weight mapping method is considerably smaller, and the nonparametric method shows slightly higher sensitivity for all tested voxel-wise thresholds except the optimal threshold of $p_{vox} = 0.05$. For the latter threshold, the parametric method achieves a 1% higher sensitivity than the nonparametric approach (see Figure 11.7 on page 108).

In the 3T tapping synchronization experiment group-level analysis, the searchlight-based nonparametric method decodes brain regions such as the motor cortex and LGN, which both are not revealed using T-based methods and the searchlight method (see Figure 11.11 on page 116). It seems very plausible that both brain regions are highly involved in encoding information for a visual tapping paradigm, because different kinds of visual synchronization stimuli are classified against each other. This not only involves the subcortical visual relay nuclei located within the LGN but it is also highly likely that the different synchronization paradigms exhibit distinguishable fingerprints in the motor cortex. Therefore, the surplus of decodable areas for the SLD method when using the nonparametric framework gives evidence for an increase in statistical sensitivity. The differences for the FWM method between nonparametric and parametric tests is smaller for the same fMRI data set. Whereas the positive weights seem to be comparable for both types of statistics, the nonparametric framework is able to identify additional plausible brain areas for negative weights: while primary visual areas remain undetected using T-based methods, the area is labeled as significant when the nonparametric framework is employed (see Figure 11.13 on page 118). As I have argued above, an involvement of the visual system in encoding information in regards to the type of visual stimulation is most highly anticipated. Hence it can be concluded that the sensitivity is indeed increased for the nonparametric statistics, regardless of the underlying information mapping technique (SLD or FWM).

13.3.2 Precision

For the group simulation it was possible to define the precision quantitatively as the ratio between voxels that were labeled as significant within the informative regions and all voxels that were found significant, hence allowing a quantitative comparison. In the case of searchlight decoding, the precision was found to be higher throughout all voxel-wise thresholds for the nonparametric framework as compared to the T-based approach (see Figure 11.3 on page 102). When using the optimal voxel-wise threshold for the SLD method ($p_{vox} = 0.001$), the precision was found to be about 6% higher (for the nonparametric framework). For the feature weight mapping method, the difference in terms of precision between the nonparametric and parametric framework turned out higher than in case of searchlight decoding. Over all voxel-wise thresholds, the nonparametric method showed higher values for the precision here (see Figure 11.7 on page 108). At the optimal voxel-wise threshold $p_{vox} = 0.05$ for the FWM method, the precision of the nonparametric framework was found to be about 13% higher than for the

parametric framework.

For the 3T tapping synchronization study, there was no clear difference visible in terms of precision between nonparametric and parametric frameworks when searchlight decoding was employed. None of the statistical frameworks labeled additional, implausible brain regions as significant (see Figure 11.11 on page 116). However, it should be noted that the nonparametric method labeled a considerably larger overall area as significant, which likely is due to a searchlight specific effect described later (see Section 14.1 on page 139). For the feature weight mapping method, the situation appears different, as the T-based methods label brain regions within white matter as significant (see Section 11.13 on page 118 in the fourth slice). As the BOLD signal from the white matter fibers is not expected to contain information about the type visual stimulation, these results most probably are false positives. Since a higher fraction of false positives implies a loss in precision, it can be stated that from a qualitative point of view, the precision of the nonparametric framework is superior to the parametric for the 3T tapping synchronization experiment using feature weight mapping.

13.3.3 Credibility

The credibility of the nonparametric framework could be measured both on the single-subject and group level using null simulations, while the parametric framework could only be employed on the group level. On the single-subject level, the null simulations were only performed for the nonparametric framework (see Section 10.2 on page 87). The results suggest that the empirically found level of false positivity for both the SLD and FWM method is within the limits, i.e. that the type I error rate is not exceeded for the nonparametric method.

On the group level, the null simulation was computed both for the nonparametric framework and T-based statistics. The level of false positivity for the nonparametric framework was within the limits for both information mapping methods (SLD and FWM). The false positive rate for T-based methods, however, was unacceptably high and drastically exceeded the expected level, both for the SLD and the FWM methods (see Section 11.2 on page 113). This suggests that type I error control is *not* ensured if T-based approaches implementing a multiple comparisons correction using Gaussian random fields are used for classification based fMRI.

13.3.4 Conclusion on the quality of nonparametric and parametric tests

In regards to the quality of the nonparametric and parametric statistical frameworks for classification-based fMRI, the nonparametric framework clearly outperforms the parametric T-based methods. This is well in line with previous research indicating an advantage to applying nonparametric statistics for univariate fMRI data analysis[121]. The advantage of nonparametric methods was found for both information mapping methods, i.e. for searchlight decoding and for feature weight mapping that I have used in my work. The results showed improvement in terms of sensitivity or precision, or, as in most cases both in sensitivity and precision

simultaneously. Furthermore, the T-based framework exhibited an extremely high degree of false positivity when tested repeatedly using a large number of null simulations, corrupting the credibility of the T-based method. The nonparametric framework, on the other hand, was found within the theoretically expected limits of false positivity.

The loss in statistical power and reliability for the T-based framework presumably reflects the previously discussed violations of the underlying formal assumptions that are violated if T-statistics are applied to group statistics for accuracy or weight maps (see [Section 13.1.1 on page 127](#)). In addition to this, the assumptions of Gaussian random field theory that are imposed on the data (e.g. sufficient smoothness) are possibly violated as well.

Chapter 14

Information mapping methods

In my thesis I present a novel nonparametric statistical framework tailored for two multivariate information mapping methods: searchlight decoding (SLD) and feature weight mapping (FWM). The underlying nonparametric frameworks for performing statistical inference are very similar for both information mapping methods. The statistical framework is based on random permutations of the order of the data points and additional resampling techniques (in the case of group studies). Notably, the same classifier algorithm including its parameters is used for both information mapping methods (SLD and FWM). As the performance and quality of the nonparametric framework has been discussed in detail in the prior Section 13.2, I want to now focus on the differences between the two information mapping techniques, by further characterizing the properties of the SLD and FWM method[122].

14.1 Searchlight decoding

The basic idea behind the searchlight approach is that the central voxel of a searchlight represents the aggregate classification result of a (commonly spherical) neighborhood of voxels[63, 5]. Repeating this over all locations yields in an information map, which is unbiased as there is no a priori spatial restriction to a subset of voxels. Importantly, the searchlight procedure implies that adjacent searchlights share a common subset of voxels, which consequently results in a spatial correlation in the accuracy maps between neighboring locations, analogously to a complex spatial filter: the spatial correlation between adjacent voxels in the accuracy maps depends on the local distribution of information and also the diameter of the searchlight, as larger searchlights share a larger subset of voxels.

Under the (credible) assumption that activity-based information is distributed relatively sparsely in the brain (i.e. mostly in the cortical surface and other grey matter locations), it appears quite intuitive that searchlight information maps are likely inflated and possibly even distorted: For instance, consider a null image containing no signal, except one single voxel containing a large amount of class information (see Figure 14.1B). The resulting searchlight information map will label practically every searchlight location which contains this voxel as

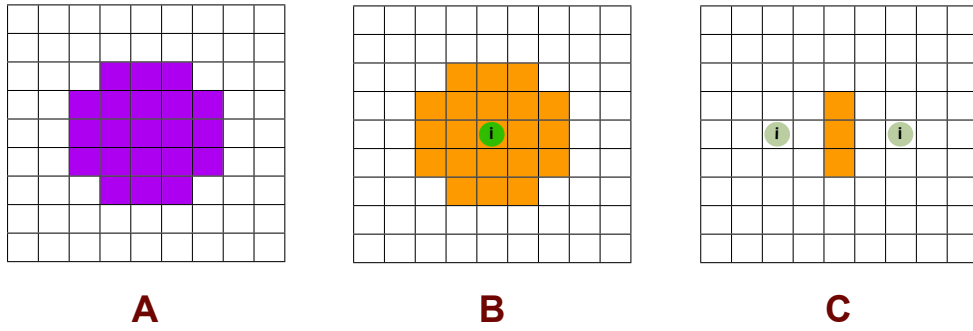


Figure 14.1: Schematic illustration of the searchlight induced inflations and distortions **(A)** Searchlight shape (down-projection to 2D) with a 5 voxel diameter (corresponding to 15mm physical size at typical 3T resolutions). The violet shaded voxels are located within the searchlight. **(B)** No voxels carry class information, except the center voxel featuring the green sphere labeled with the letter “i”: this voxel is the sole voxel carrying class information. As a result of the SLD procedure using the searchlight of **A**, many voxels are being labeled as informative (these voxels are depicted in orange). The inflating effect has previously been termed as “needle in the haystack effect” [123] **(C)** Here, no voxels except the two voxels with the green sphere labeled with “i” carry class information. The information carried by one voxel, however, is sufficiently small so that a searchlight has to include *both* informative voxels in order to be labeled significant. Hence only the voxels in the middle, where the searchlight contains both informative voxels are labeled informative, resulting in inaccurate and distorted information maps.

informative, effectively grossly inflating the actual informative regions. In a recent study, this effect had been termed as “*needle in the haystack effect*” [123]. The reverse of this effect is also conceivable, where the center voxel of a searchlight *does not* contain any information, while information is solely present at the searchlight’s edge. Labeling the center voxel significant thus results in a distortion (see Figure 14.1C); conveniently this effect had been given the name “*haystack in the needle*” [123]. Furthermore, it has been shown that the latter effect depends on the searchlight diameter, as the number of informative voxels monotonically depends on the diameter of the searchlight. However, the effect also depends on the distribution of information and overall geometry. Lastly, the (multivariate) signal to noise ratio presumably also plays an important role.

In this regard, it appears questionable whether the searchlight is the optimal weapon of choice when it comes to maximally exploiting the benefits of high-field or even ultra-high-field fMRI[45]. From an empirical point of view, the above considerations in regards to inflation and distortion are fully supported: the searchlight method indeed yields inflated estimates of information distribution in both the simulated and real fMRI data sets. This effect becomes especially visible for high-resolution data sets (7T finger tapping and imagination, see Figure 10.12A), where there is a high fraction of significant voxels *outside* of grey matter areas, obscuring the actual distribution of information in the cortex. Also in simulations where the information is distributed in fine spatial scales (the single subject geometric simulation, see Figure 10.1 on page 82), the spatial inaccuracies of the resulting information maps are clearly

visible. The effect of inflation and distortion is also noticeable at lower fMRI resolutions and group studies (see Figure 11.15 on page 120). Especially problematic is the case where there exist two adjacent regions, which each encode class information of different classes. In here, the searchlights centered around the border between both of these regions are possibly the most informative ones. This effect likely took place in the group 3T tapping synchronization experiment (see Figure 11.15, first slice), where the two differently coding regions (primary visual and visual association area) were identified by the FWM method. The searchlight method assigned the highest decoding accuracies at the border of both regions. Similarly, the effect could also be reproduced in the single-subject simulation, where the border areas received (on average) the highest accuracies (see Figure 10.1 on page 82).

The inflation effect clearly depends on the searchlight's diameter, as larger searchlight diameters also increase the sensitivity in general terms. (however, this gain also depends on the distribution of information). On the other hand, larger searchlight diameters decrease the precision obtained (see Figure 11.4 on page 104). A further effect of larger searchlight diameters is an increase in spatial correlation in the resulting searchlight maps. This is the case because the volume of overlap between adjacent searchlights, i.e. the number of features that are shared across two neighboring searchlight locations, monotonically depends on the diameter of the searchlights. As the nonparametric method implicitly considers the underlying smoothness of the images (see Figure 10.6 on page 88), the resulting in *broader* empirical cluster histograms may obscure ground truth clusters.

The issue of inflation may be mitigated by limiting the searchlight only to grey matter voxels or even directly applying it on the cortical surface[1]. However I want to point out that the reduction of inflation is only constrained to one of three spatial dimensions for the surface-based methods; while the spatial accuracy in the direction normal to the cortical surface is improved, the two in-plane dimensions (along the cortical sheet) remain inflated and distorted. Furthermore, it should be mentioned that in the case of the grey matter masking, it is crucially necessary to find a mask that is valid for the entire group of subjects for performing voxel-wise group statistics, since all subjects need to have accuracy values for a given voxel to properly compute statistics.

Another aspect worth discussing is the property of *localness* of the searchlight method. I want to stress that the searchlight approach only analyzes a *small neighborhood* of voxels at a time. However, given that the brain is a large and heavily interconnected network, it appears very plausible that the fingerprint of distinct brain states does not *solely* exist at *small* spatial scales, i.e. from local processing. More likely, the brain processes information on larger spatial dimensions across distinct networks. For instance, remote brain areas do jointly exhibit patterns of activation governed by long-range neural communication[124]. Evidently, such large-scale interactions cannot be captured by the searchlight method. However, capturing large-scale information patterns is not always the goal of imaging studies, as sometimes a strict locality and investigation of small spatial scales is desired. An example for such a case is a study where finger movement is decoded from within the *ipsilateral*¹ cortical hemisphere[125].

¹on the same side of the body as the hand is, e.g. both on the left body side

14.2 Feature weight mapping

As in the case for the searchlight method, the feature weight mapping approach is spatially unbiased, since the whole brain is analyzed without prior restriction to a specific spatial hypothesis (e.g. region of interest). In contrast to the searchlight method, however, the FWM approach uses the information of all voxels *simultaneously* as input. In the implementation used in my thesis, a PCA-based procedure for the reduction of the dimensionality of the data is applied beforehand in order to decrease the number of degrees of freedom for the classifier's parameters. A second mark of distinction to the SLD method is that no cross-validation is carried out in the FWM method, the classifier is only trained once on the entire data set of a single subject. This results in a weight vector for the classifier's training, which is mapped back from the PCA space into the voxel space. Using the nonparametric framework, the statistical inference is then carried out on the weight vector acquired from the classifier's training in the voxel space[122].

Since the classification has to be computed only a single time on a relatively small data set (for each permutation), the computational resources necessary for the nonparametric statistical framework are drastically lower than compared to the ones needed for searchlight decoding (as in the SLD method classification is computed for every location and every cross-validation step). Depending on the size of the data set in terms of voxels (resolution) and experimental trials, the computation of the permutations in the FWM method is between 5000 and 30000 times faster than the SLD method. For instance, the average computation time for one permutation in the group simulation 5cubes was 0.008s for the FWM method, while the same computation took 38s for the SLD method (carried out on a single CPU). For the ultra-high resolution data set (7T finger tapping and imagination), the difference in computation speed was even larger; one permutation was computed in 0.0057s for the FWM method while one permutation took 167s for the SLD method.

A further distinction between the FWM and SLD method is that in the FWM method each feature dimension (i.e. voxel or principal component) is assigned a weight as a result of the classifier's training (in contrast to decoding accuracies for the SLD method). The weight itself is directly derived from the underlying classifier's mathematical model (in the case of my thesis, a linear support vector machine) and is an indicator of the *contribution* of the corresponding feature dimension to the classification decision. For a linear classifier, the predicted class of an unseen data point \vec{y} is determined by the sign of the dot product between the weight vector and the data point $\vec{w} \cdot \vec{y} = w_1y_1 + w_2y_2 + \dots + w_{N_{vox}}y_{N_{vox}}$. Thus, the size of the weight vector at the k -th dimension w_k can be interpreted as the *importance* of the k -th feature dimension (the k -th voxel). In other words, the absolute size of the component w_k directly translates into a high involvement in the class prediction of the corresponding feature dimension k . It should be highlighted that a weight component w_k is either of *positive* or *negative* algebraic sign. The sign of the component indicates to which class the k -th feature or voxel *influences* the classification decision; for instance if a voxel activates consistently when in class A but does not activate when in class B, the resulting weight component would be positive. On the other hand, if the voxel activates consistently when in class A but does not activate when in class B, the weight component would be determined negative. Hence, the individual weight component

reveals how the corresponding feature dimension (voxel) influences the classification decision depending on the level of activation found in the feature. In an area with positive weights, a high activity level would have an impact for the classifier to decide for class A, while a low level of activity favors the decision for class B. For negative weights, an analogous argument can be made; in here a high level of activity would influence the classifier to decide for class B and a low level of activity brings a favor for class A. In contrast, the SLD method is not able to deliver such information about the *direction* of influence for any given features, as the SLD method only determines *whether* class information is present or not.

The directionality of the weight components effectively increases the *interpretability* of the data. This becomes most clear in the group-level analysis of the 3T tapping synchronization experiment (see Figure 11.15 on page 120), where two visual tapping conditions (visual discrete and visual continuous) are classified against each other. The FWM method depicts the primary visual cortex with positive weights and secondary visual regions (Brodmann area 18) with negative weights. This implies that large activations in the primary visual system influence the classifier to decide for *continuous* visual stimulation, while large activations in the visual association areas favor a decision for *discrete* visual stimulation. It should be noted, however, that the reverse is also true, i.e. low activations in primary regions favor the classifier for discrete visual stimulations and low activations in the visual association areas favor a decision for continuous visual stimulation. On the other hand, the searchlight method solely is able to map out the visual cortex being discriminative for the two conditions, with especially high accuracies at the class borders derived from the FWM method.

Another aspect in regards to interpretability is that a weight component in the FWM method located at voxel k does *solely* represent the influence to the classification of the k -th voxel. As a result, information maps with a very high spatial accuracy can be obtained. This is especially advantageous for single-subject studies, where due to the lack of anatomical variability (which is present in between a group of subjects) a very precise localization of informative regions is principally possible. Indeed, for the single-subject geometric simulation (see Figure 10.1 on page 82), the FWM method precisely delineates the informative regions. Also in the case of the ultra-high 7T finger tapping and imagination data (see Figure 10.12 on page 95), only the surface of the cortex is labeled as informative. Since white matter fibers, which do not contain task-information in their BOLD signal are located inside cortical regions, the FWM results are rendered very credible. Hence the FWM method is especially advantageous for maximally exploiting higher resolutions in the sense that the mapping local information content is reliable and precise.

In contrast, in the searchlight decoding method the accuracy at voxel k characterizes the *aggregate* decodability of a *neighborhood* of voxels around the k -th voxel (typically this neighborhood is spherical). In other words, for the SLD analysis technique, voxels with high accuracies do not necessarily have to be informative; the informative voxels may be located elsewhere in the voxel's neighborhood. This aspect is problematic, as this important distinction commonly is ignored in the neuroscientific practice where searchlight-based analysis are employed[126].

14.3 Conclusion on information mapping techniques

To summarize, the main difference between the two information mapping methods used in my work (namely searchlight decoding and feature weight mapping) in terms of empirical results is their *spatial precision*. The searchlight decoding method returns inflated and distorted results and has the inherent limitation that a voxel of a searchlight accuracy map does not represent the information content of this voxel but rather the information content of the neighborhood of voxels around this voxel. In contrast to this, a voxel in a weight map computed by the feature weight mapping method does represent the influence of each voxel exclusively to the classification decision. Consequently, the FWM method returns very precise information maps in practice[122].

The statistical power of both methods critically depends on the underlying geometry of the distribution of information: if the information is distributed in a concentrated fashion at coarse scale (e.g. cubes of information, see Figure 10.5 on page 85), the searchlight decoding method is favored. In the case of a fine information distribution (e.g. cortical layers), the feature weight mapping method is favored (see see Figure 10.5). Ultimately, the distribution of information may be tend towards a fine distribution; this is especially true for higher resolutions and single subject studies, where it is possible to resolve the structure of the cortical layering.

In terms of computational time, the picture is very clear: in the simulations and fMRI data sets used for this thesis, the gain in speed was between 5000 and 30000 times for the FWM method as compared to searchlight decoding.

Chapter 15

Single subject or group studies

In my thesis I have applied the new proposed nonparametric statistical framework on two levels on inference: on the level of *single-subject* studies and on the *group level*. Both levels of inference compromise distinct motivations and imply different assumptions. Consequently, the scientific statements based on the two inference methods are also differing. In the following, I want to discuss the rationale behind each level of inference and their underlying assumption from a general neuroscientific point of view.

15.1 Motivation for group level inference

Most commonly, experiments in imaging-based cognitive neuroscience are carried out on the group level. The main theoretical motivation to study brain function on a group level is the fact that it is desirable to identify brain processes that are *universal* within a population. This implies the assumption that there exist universal spatio-temporal aspects of brain dynamics that are shared amongst a group of subjects. The assumption of universality is compounded by the further assumptions, e.g. that deviations from the universal “fingerprint” of these brain processes are due to the noisy character of brain function.

The above approach has been proven to enable profound insights into body function from a physiological point of view. Consider for instance the study of any internal organ of the human body: Following the above assumptions, it is possible to *abstract* general features of functionality that are universal amongst the population of human beings. Using these abstracted features, the role of the organ and all its constituents can be understood in a general sense. Furthermore, it is possible to clearly delineate inter-individual differences, which may themselves be linked to genetic or environmental factors. As the principle of universal function has been extremely fruitful for vast fields of biology and in particular physiology, it seems reasonable to conclude that the same principles can be applied to study the organ of ultimate interest and highest complexity: the human brain.

From a more pragmatic point of view, the power of statistical tests is higher on a group level as compared to single-subject studies, especially if the number of included subjects

is large[121]. This relation also holds for the case for pattern classification, as I have shown that the sensitivity of the group simulation 5cubes result depends on the number of included subjects (see Figure 11.10 on page 111). In regards to fMRI data, the results of the 3T tapping synchronization experiment on the group level (see Figure 11.15 on page 120) are much richer as compared to the single-subject analysis (Figure 10.7 on page 89). The reason for the gain in statistical power on a group level as compared to the single-subject studies is the high experimental variance. Critically, this variance is considered as noise and it is assumed that through averaging, the effects of this noise can be removed[127].

15.2 Motivation for the analysis on the single-subject level

With the advent of more sensitive analysis methods, higher magnetic field strengths and more advanced study designs¹it has become clear that group level studies are not the optimal basis for an adequate description of brain dynamics. This is mainly due to the fact that the variations across subjects are comparably large across subjects, resulting in group maps that are not any more representative for the individual scale[127]. In other words, group level inference methods depict only the smallest common denominator of the local activation or information content; only if a sufficiently large number of subjects features brain activity at a given location, the group map at this location will include the effect. If an activation is shared only within a small subset of the entire group, the activation is considered as noise. Most crucially, however, it was shown that the single subject spatial patterns of activity are stable over scanning session, i.e. reproducible [128, 129]. The inter-individual differences arise from a variety of factors, in particular the cognitive style and strategy: for instance, it has been shown that subjects with similar cognitive strategies in regards to brain tasks exhibit a greater number of similar brain activation patterns [130]. Furthermore, it should be noted that the variability is not only found on functional side, but also in terms of brain anatomy[98].

Conclusively, it can be stated that it is premature to consider inter-subject variations as a manifestation of noise, as there is compelling evidence showing that these variations are indeed related to the brain function on the scale of an individual level. The stability of the variations rather supports the hypotheses that *every* brain is used *differently* and that claims about universal brain processes, in particular for high-level cognition, should be looked upon rather critically.

Studying brain imaging data on the level of single subjects resolves many of the above issues. Ultimately, single-subject analysis even allows one to draw conclusions on the population of subjects, as it can be investigated which individual brain activity patterns are *normative* for a group[127]. Seen from another perspective, single subject analysis allow insights into the *source* of the variability.

Another aspect worth considering is that from a principle point of view, single-subject analysis makes it possible to also *categorize* subjects on basis of their spatio-temporal patterns of brain activity into distinct groups. For instance, a possible categorization can be done in

¹such as systematic test-retest studies

correspondence to certain cognitive strategies or even risk groups for brain diseases such as schizophrenia[131].

In regards to the data analysis on the single-subject level, the otherwise indispensable preprocessing step of spatial normalization is not strictly necessary². The main advantage of this is the possibility of an extremely precise localization of activation patterns. This gain of spatial specificity ultimately makes it possible to infer highly precise relations between brain structure and brain function, which are not possible on a group level due to inter-subject anatomical variability: the spatial normalization into a common brain space effectively implies a loss in resolution, which is in the range of centimeters. Hence, for ultra-high field studies using high resolutions, the disadvantages of standard spatial normalization undermine the gain in resolution, rendering the analysis on the single-subject level as the only feasible choice.

15.3 Conclusion on level of inference

In conclusion it can be stated that studies on a level of the single subject are likely going to play a larger role in the future of imaging-based cognitive neuroscience, as they principally allow insights that are outside the range of group inference. However, progress in terms of sensitivity and power for the image acquisition, analysis and modeling methods are a condition *sine qua non* for this. For instance it is possible that the average scanning times used in neuroscientific experiments may not be sufficient, as an increase in scanning time beyond one hour results in a monotonic gain of statistical power[132].

²the spatial normalization is not necessary unless conclusions on group level based on single-subject analysis are of special interest. This was the case for the 3T tapping synchronization experiment, the single subject analysis took place in the MNI standard space to ensure comparability to the group results.

Part V

Conclusion

Chapter 16

Summary

In the following, I want to briefly summarize the most important findings and conclusions of my thesis. For the sake of clarity I present the summary in bullet point form.

I have introduced a novel nonparametric statistical framework for classification-based fMRI, which is based on random permutations (on the single-subject level) and random permutations in combination with resampling techniques (on the group level). In both cases, cluster-based statistics are employed for multiple-comparisons correction. The framework hereby relies on minimal assumptions. The main results and conclusions are:

- the nonparametric framework is applicable for two distinct information mapping methods (searchlight decoding and feature weight mapping)
- the framework makes it possible to infer statistical significance on the single-subject level
- on the group level, I have extensively compared the nonparametric framework versus commonly practiced T-based statistics. Here, the main findings were:
 - the nonparametric framework has a higher sensitivity than T-based methods
 - the spatial precision is higher for the nonparametric method
 - the credibility (in terms of false positivity) is on an adequate level for the nonparametric framework, however, for T-based methods the credibility is substantially inferior
 - T-based methods are inappropriate from a theoretical point of view if applied to classification-based fMRI; the same holds for the parametric binomial model if cross-validations are applied

The two information mapping methods, namely searchlight decoding and feature weight mapping, were compared extensively. The main results and conclusions were:

- In order to achieve a comparable ratio between the sensitivity and precision of both information mapping methods, the voxel-wise thresholds for searchlight decoding has to be set considerably higher (i.e. lower p -values)
- Searchlight decoding systematically produces inflated and distorted results
- In scientific practice, searchlight accuracy maps should be interpreted with caution, as a significant voxel does not imply that this voxel indeed carries information (it implies that the *neighborhood* of voxels contains information)
- The feature weight mapping method returns information maps of a very high spatial precision. This especially holds if the distribution of information is rather fine, which is the case for ultra-high resolution fMRI
- The computation time needed for the feature weight mapping method is tiny compared to the time needed to perform searchlight decoding

16.1 Limitations

While I have shown that the proposed nonparametric framework clearly outperforms T-based statistics for carrying out statistical inference, the proposed framework also exhibits certain limitations. In the following, I want to discuss the main limitations given my point of view. It should be noted that for some of these limitation, I provide possible solutions in the next Section [16.2](#).

In practical terms, the biggest limitation is the choice of the free parameter p_{vox} . This parameter specifies how unlikely a decoding accuracy or feature weight needs to be in order to be considered as candidate for an informative region. There is no first principle deduction of this value, however by convention values of $p_{vox} > 0.05$ are generally considered inadequate. The results of the nonparametric framework highly depends on the choice of this threshold. If the threshold is set too high (i.e. low p -values), the sensitivity of the tests becomes too small. At the same time, the spatial precision becomes better for higher thresholds. Notably, the same relation holds for parametric frameworks, such as the T-based method. More generally, this trade-off between sensitivity and precision is inherent to frequentist statistics and depends on the free parameters of the test (see Figure [5.1 on page 33](#)).

A further limitation of the current nonparametric framework is that the spatial correlation (i.e. smoothness) of the underlying images is considered implicitly and influences the empirical cluster statistics on a global scale, as the underlying smoothness has a *global* effect on the statistics applying on the entire volume. Local deviations in terms of smoothness are not directly considered: consider for instance the extreme case of two areas of similar size, one with a very small amount of spatial correlation and the other with a large smoothness. Both areas will have equal impact on the empirical cluster statistics, however if considered separately, the empirical cluster size records would differ. Effectively, due to the above considerations in regards to the global impact, it is possible that in this scenario higher levels of false negativity

within the area of low smoothness may occur. In the area of higher smoothness, however, higher levels of false positivity could theoretically occur. On the other hand, it should be noted that the T-based corrections in SPM8 used in my work also consider the smoothness only on a global level.

Another limitation is the inherent assumption of the proposed framework, that information is distributed in clusters of voxels as opposed to information within very small spatial scales (smaller than one voxel). With the current framework, such fine-distributed information cannot be captured. It should be noted, however, that the cluster assumption to some degree is undermined in the spatial correlations inherent in the data, which potentially dilutes very small spatial scales of brain activation.

Finally, I want to stress a more general point regarding the usage of machine learning approaches in neuroimaging, namely the *black-box* problem. As I have already stated in Section 4.4.6.1 on page 28, information mapping methods allow the delineation of brain regions which contain information about specific brain states that are for instance provoked by different stimulus conditions. While information mapping methods make it possible to know *where* brain-state specific information is present in the brain, the machine learning approaches used in my work do not allow any further insight into *how* these brain-states are realized on a deeper biophysical level. However, a quantitative description of the underlying biophysical processes in the form of a model is ultimately one of the most important goals of neuroscience, if not the *most* important¹. The information mapping approaches do not offer any insights into these research questions and possibly even obscure a more biophysically motivated point of view.

16.2 Outlook

In the following I want to provide the most promising extensions to the nonparametric framework, that might be applied to the groundwork introduced in this thesis.

An interesting extension would be an adaptation of the threshold-free cluster enhancement techniques [120] tailored for classification-based fMRI. Effectively, an adaptation would render the choice of the voxel-wise threshold obsolete and substitute this choice by a set of parameters, which have a smaller impact on the resulting test statistics. However, as discussed above, the derivation of this substitution suitable for general classification-based fMRI data may not be trivial.

Another point is that the main computational bottleneck is the computation of the permutations, followed by the cluster-based statistics. Since these computations are not overly memory intensive and can be parallelized to a high degree, the calculation can be performed on modern graphics processing units (GPUs). This results in a vast reduction of computation time, especially for the searchlight decoding approach. Furthermore, GPU methods may also accelerate the processing for the feature weight mapping method, eventually making a full inference within the timeframe of seconds possible.

¹given the author's point of view

Building on an availability of increased processing power, it also would be possible to fully integrate differences in local spatial correlation, which in the proposed current framework are only considered on a global level. By carrying out the empirical cluster statistics for each location (yielding location-specific cluster size histograms), it would be possible to further increase both the sensitivity and precision of the nonparametric framework. Furthermore, using voxel-specific cluster statistics it would be possible to nonparametrically derive a measure for the local spatial correlations.

References

- [1] Y. Chen, P. Namburi, L. T. Elliott, J. Heinzle, C.-S. Soon, M. W. L. Chee, J.-D. Haynes, Cortical surface-based searchlight decoding., *Neuroimage* 56 (2) (2011) 582–592.
- [2] T. E. Nichols, A. P. Holmes, Nonparametric permutation tests for functional neuroimaging: a primer with examples, *Human Brain Mapping* 15 (1) (2002) 1–25.
- [3] P. Golland, F. Liang, S. Mukherjee, D. Panchenko, Permutation Tests for Classification, , (2005) 330–341.
- [4] J. Mourão-Miranda, A. L. W. Bokde, C. Born, H. Hampel, M. Stetter, Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data, *Neuroimage* 28 (4) (2005) 980–995.
- [5] J. Stelzer, Y. Chen, R. Turner, Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control., *Neuroimage* 65 (2013) 69–82.
- [6] Lauterbur, P.C., Image formation by induced local interactions: examples employing nuclear magnetic resonance, *Nature* 242 (5394) (1973) 190–191.
- [7] P. Suetens, *Fundamentals of Medical Imaging*, Cambridge Univ Press, 2009.
- [8] F. Bloch, W. W. Hansen, Nuclear induction, *Physical review* 70 (1946) 460–474.
- [9] M. H. Levitt, *Spin dynamics, basics of nuclear magnetic resonance*, Wiley, 2008.
- [10] W. Gerlach, O. Stern, Der experimentelle Nachweis der Richtungsquantelung im Magnetfeld., *Zeitschrift für Physik A Hadrons and Nuclei* 9 (1922) 1.
- [11] P. W. Stroman, *Essentials of Functional MRI*, CRC Press, 2011.
- [12] D. W. McRobbie, E. A. Moore, M. J. Graves, *MRI from Picture to Proton*, 2nd Edition, Cambridge Univ Press, 2007.
- [13] L. Pauling, C. D. Coryell, *The Magnetic Properties and Structure of Hemoglobin*, Oxy-

hemoglobin and Carbonmonoxyhemoglobin, Proceedings of the National Academy of Sciences of the United States of America 22 (4) (1936) 210.

- [14] K. R. Thulborn, J. C. Waterton, P. M. Matthews, G. K. Radda, Oxygenation dependence of the transverse relaxation time of water protons in whole blood at high field, *Biochimica et Biophysica Acta (BBA)-General Subjects* 714 (2) (1982) 265–270.
- [15] S. Ogawa, T. M. Lee, A. R. Kay, D. W. Tank, Brain magnetic resonance imaging with contrast dependent on blood oxygenation., *Proceedings of the National Academy of Sciences of the United States of America* 87 (24) (1990) 9868–9872.
- [16] C. S. Roy, C. S. Sherrington, On the Regulation of the Blood-supply of the Brain., *The Journal of physiology* 11 (1-2) (1890) 85–158.17.
- [17] N. K. Logothetis, What we can do and what we cannot do with fMRI, *Nature* 453 (7197) (2008) 869–878.
- [18] N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, A. Oeltermann, Neurophysiological investigation of the basis of the fMRI signal, *Nature* 412 (6843) (2001) 150–157.
- [19] P. T. Fox, M. E. Raichle, Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during somatosensory stimulation in human subjects, *Proceedings of the National Academy of Sciences* 83 (1986) 1140–1144.
- [20] N. K. Logothetis, B. A. Wandell, Interpreting the BOLD Signal, *Annual Review of Physiology* 66 (1) (2004) 735–769.
- [21] P. Fox, M. Raichle, M. Mintun, C. Dence, Nonoxidative glucose consumption during focal physiologic neural activity, *science* 241 (4864) (1988) 462–464.
- [22] P. J. Magistretti, L. Pellerin, Cellular mechanisms of brain energy metabolism and their relevance to functional brain imaging, *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 354 (1387) (1999) 1155–1163.
- [23] R. B. Buxton, L. R. Frank, A Model for the Coupling Between Cerebral Blood Flow and Oxygen Metabolism During Neural Stimulation., *Journal of cerebral blood flow and metabolism* 17 (1) (1997) 64–72.
- [24] J. H. Lee, R. Durand, V. Gradinaru, F. Zhang, I. Goshen, D.-S. Kim, L. E. Fenno, C. Ramakrishnan, K. Deisseroth, Global and local fMRI signals driven by neurons defined optogenetically by type and wiring, *Nature* 465 (7299) (2010) 788–792.
- [25] S. Moeller, E. Yacoub, C. A. Olman, E. Auerbach, J. Strupp, N. Harel, K. Ugurbil, Multi-band multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI., *Magnetic resonance in medicine* 63 (5) (2010) 1144–1153.

- [26] M. Lustig, D. Donoho, J. M. Pauly, Sparse MRI: The application of compressed sensing for rapid MR imaging, *Magnetic resonance in medicine* 58 (6) (2007) 1182–1195.
- [27] N. Kriegeskorte, R. Cusack, P. Bandettini, How does an fMRI voxel sample the neuronal activity pattern: compact-kernel or complex spatiotemporal filter?, *Neuroimage* 49 (3) (2010) 1965–1976.
- [28] P. L. Purdon, R. M. Weisskoff, Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates in fMRI., *Human Brain Mapping* 6 (4) (1998) 239–249.
- [29] Spm8 wellcome trust centre for neuroimaging, london, -Software available at <http://www.fil.ion.ucl.ac.uk/spm/software/spm8>.
- [30] M. Jenkinson, C. F. Beckmann, T. E. J. Behrens, M. W. Woolrich, S. M. Smith, FSL., *Neuroimage* 62 (2) (2012) 782–790.
- [31] R. W. Cox, AFNI: software for analysis and visualization of functional magnetic resonance neuroimages., *Computers and biomedical research, an international journal* 29 (3) (1996) 162–173.
- [32] G. Lohmann, J. Stelzer, J. Neumann, R. Turner, N. Ay, “More is different” in fMRI: a review of recent data analysis techniques, *Brain Connectivity* (2012) 1–30.
- [33] J.-B. Poline, M. Brett, The general linear model and fMRI: Does love last forever?, *Neuroimage* (2012) 1–10.
- [34] A. M. Wink, J. B. T. M. Roerdink, BOLD Noise Assumptions in fMRI., *International Journal of Biomedical Imaging* 2006 (2006) 12014.
- [35] K. J. Friston, A. P. Holmes, K. J. Worsley, J. P. Poline, C. D. Frith, R. S. Frackowiak, Statistical parametric maps in functional imaging: a general linear approach, *Human Brain Mapping* 2 (4) (1994) 189–210.
- [36] G. Lohmann, S. Hoehl, J. Brauer, C. Danielmeier, I. Bornkessel-Schlesewsky, J. Bahlmann, R. Turner, A. Friederici, Setting the frame: the human brain activates a basic low-frequency network for language processing., *Cerebral cortex* 20 (6) (2010) 1286–1292.
- [37] J. R. Moeller, S. C. Strother, A regional covariance approach to the analysis of functional patterns in positron emission tomographic data., *Journal of cerebral blood flow and metabolism* 11 (2) (1991) A121–35.
- [38] C. M. Clark, W. Ammann, W. R. Martin, P. Ty, M. R. Hayden, The FDG/PET methodology for early detection of disease onset: a statistical model., *Journal of cerebral blood flow and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism* 11 (2) (1991) A96–102.

- [39] A. R. McIntosh, F. L. Bookstein, J. V. Haxby, C. L. Grady, Spatial pattern analysis of functional brain images using partial least squares., *Neuroimage* 3 (3 Pt 1) (1996) 143–157.
- [40] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, P. Pietrini, Distributed and overlapping representations of faces and objects in ventral temporal cortex., *science* 293 (5539) (2001) 2425–2430.
- [41] D. D. Cox, R. L. Savoy, Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex., *Neuroimage* 19 (2 Pt 1) (2003) 261–270.
- [42] T. A. Carlson, P. Schrater, S. He, Patterns of Activity in the Categorical Representations of Objects, *Journal of Cognitive Neuroscience* 15 (5) (2003) 704–717.
- [43] C.-S. Soon, M. Brass, H.-J. Heinze, J.-D. Haynes, Unconscious determinants of free decisions in the human brain., *Nature Neuroscience* 11 (5) (2008) 543–545.
- [44] N. Kriegeskorte, Pattern-information analysis: from stimulus decoding to computational-model testing., *Neuroimage* 56 (2) (2011) 411–421.
- [45] N. Kriegeskorte, P. Bandettini, Analyzing for information, not activation, to exploit high-resolution fMRI., *Neuroimage* 38 (4) (2007) 649–662.
- [46] J. Langford, Tutorial on Practical Prediction Theory for Classification, *Journal of Machine Learning Research* 6 (6) (2005) 273–306.
- [47] F. Pereira, M. Botvinick, Information mapping with pattern classifiers: a comparative study, *Neuroimage* 56 (2) (2011) 476–496.
- [48] M. Misaki, Y. Kim, P. A. Bandettini, N. Kriegeskorte, Comparison of multivariate classifiers and response normalizations for pattern-information fMRI, *Neuroimage* 53 (1) (2010) 103–118.
- [49] A. J. O’Toole, F. Jiang, H. Abdi, N. Pénard, J. P. Dunlop, M. A. Parent, Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data, *Journal of Cognitive Neuroscience* 19 (11) (2007) 1735–1752.
- [50] M. Yousef, S. Jung, L. C. Showe, M. K. Showe, Recursive Cluster Elimination (RCE) for classification and feature selection from gene expression data, *BMC bioinformatics* 8 (1) (2007) 144.
- [51] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, *The Journal of Machine Learning Research* 3 (2003) 1157–1182.
- [52] F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, E. Formisano, Combin-

- ing multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns, *Neuroimage* 43 (1) (2008) 44–58.
- [53] D. E. Goldberg, J. H. Holland, *Genetic Algorithms and Machine Learning*, Machine Learning (1988) 95–99.
 - [54] O. Boehm, D. R. Hardoon, L. M. Manevitz, Classifying cognitive states of brain activity via one-class neural networks with feature selection by genetic algorithms, *Int. J. Mach. Learn. & Cyber.* (2011) 125–134.
 - [55] R. A. Poldrack, Region of interest analysis for fMRI., *Social Cognitive and Affective Neuroscience* 2 (1) (2007) 67–70.
 - [56] J.-D. Haynes, G. Rees, Predicting the orientation of invisible stimuli from activity in human primary visual cortex., *Nature Neuroscience* 8 (5) (2005) 686–691.
 - [57] Y. Kamitani, F. Tong, Decoding the visual and subjective contents of the human brain, *Nature Neuroscience* 8 (5) (2005) 679–685.
 - [58] K. A. Norman, S. M. Polyn, G. J. Detre, J. V. Haxby, Beyond mind-reading: multi-voxel pattern analysis of fMRI data., *Trends in Cognitive Sciences* 10 (9) (2006) 424–430.
 - [59] S. B. Eickhoff, K. E. Stephan, H. Mohlberg, C. Grefkes, G. R. Fink, K. Amunts, K. Zilles, A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data., *Neuroimage* 25 (4) (2005) 1325–1335.
 - [60] J. A. Etzel, V. Gazzola, C. Keysers, An introduction to anatomical ROI-based fMRI classification analysis., *Brain research* 1282 (2009) 114–125.
 - [61] S. Geyer, M. Weiss, K. Reimann, G. Lohmann, R. Turner, Microstructural Parcellation of the Human Cerebral Cortex - From Brodmann’s Post-Mortem Map to in vivo Mapping with High-Field Magnetic Resonance Imaging., *Frontiers in Human Neuroscience* 5 (2011) 19.
 - [62] S. C. Strother, J. Anderson, L. K. Hansen, U. Kjems, R. Kustra, J. Sidtis, S. Frutiger, S. Muley, S. LaConte, D. Rottenberg, The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework., *Neuroimage* 15 (4) (2002) 747–771.
 - [63] N. Kriegeskorte, R. Goebel, P. Bandettini, Information-based functional brain mapping, *Proceedings of the National Academy of Sciences of the United States of America* 103 (10) (2006) 3863–3868.
 - [64] Duda, *Pattern classification*, 2nd Edition, John Wiley & Sons, 2007.
 - [65] M. K. Carroll, G. A. Cecchi, I. Rish, R. Garg, A. R. Rao, Prediction and interpretation of distributed neural activity with sparse models., *Neuroimage* 44 (1) (2009) 112–122.

- [66] R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society Series B (Methodological)* 58 (1996) 267–288.
- [67] A. E. Hoerl, R. W. Kennard, Ridge Regression: Applications to Nonorthogonal Problems, *Technometrics* 12 (1) (1970) 69–82.
- [68] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2) (2005) 301–320.
- [69] S. Ryali, K. Supekar, D. A. Abrams, V. Menon, Sparse logistic regression for whole-brain classification of fMRI data, *Neuroimage* 51 (2) (2010) 752–764.
- [70] E. Formisano, F. De Martino, G. Valente, Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning, *Magnetic Resonance Imaging* 26 (7) (2008) 921–934.
- [71] T. Naselaris, K. N. Kay, S. Nishimoto, J. L. Gallant, Encoding and decoding in fMRI., *Neuroimage* 56 (2) (2011) 400–410.
- [72] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, J. L. Gallant, Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies, *Current Biology* 21 (19) (2011) 1641–1646.
- [73] D. G. Mayo, D. R. Cox, Frequentist statistics as a theory of inductive inference, *Lecture Notes-Monograph Series* (2006) 77–97.
- [74] P. I. Good, *Permutation, Parametric and Bootstrap Tests of Hypotheses*, 3rd Edition, Springer Science+Business Media, 2005.
- [75] W. T. Shaw, Sampling Student’s T distribution-use of the inverse cumulative distribution function, *Journal of Computational Finance* 9 (4) (2006) 37.
- [76] S. S. Shapiro, *An Analysis of Variance Test for Normality (complete Samples)* (1964) 591.
- [77] N. M. Razali, Y. B. Wah, Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests, *Journal of Statistical Modeling and Analytics* 2 (1) (2011) 21–33.
- [78] P. I. Good, *Resampling methods: A practical guide to data analysis*, 3rd Edition, Birkhauser, 2006.
- [79] A. P. Holmes, R. C. Blair, J. D. Watson, I. Ford, Nonparametric analysis of statistic images from functional mapping experiments, *Journal of cerebral blood flow and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism* 16 (1) (1996) 7–22.

- [80] C. R. Genovese, N. A. Lazar, T. Nichols, Thresholding of statistical maps in functional neuroimaging using the false discovery rate., *Neuroimage* 15 (4) (2002) 870–878.
- [81] Y. Benjamini, Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society. Series B ...* 57 (1995) 289–300.
- [82] Y. Benjamini, A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence, *Journal of Statistical Planning and Inference* (1999) 163–170.
- [83] J. R. Chumbley, K. J. Friston, False discovery rate revisited: FDR and topological inference using Gaussian random fields, *Neuroimage* 44 (1) (2009) 62–70.
- [84] S. Hayasaka, T. E. Nichols, Validating cluster size inference: random field and permutation methods, *Neuroimage* 20 (4) (2003) 2343–2356.
- [85] S. Hayasaka, K. L. Phan, I. Liberzon, K. J. Worsley, T. E. Nichols, Nonstationary cluster-size inference with random field and permutation methods, *Neuroimage* 22 (2) (2004) 676–687.
- [86] A. M. Hasofer, JSTOR: Advances in Applied Probability, *Advances in Applied Probability* 10 (1978) 14–21.
- [87] K. Worsley, Local maxima and the expected euler characteristic of excursion sets of chi-squared, F and t fields, *Advances in Applied Probability* (1994) 13–42.
- [88] E. Bullmore, C. Long, J. Suckling, J. Fadili, G. Calvert, F. Zelaya, T. A. Carpenter, M. Brammer, Colored noise and computational inference in neurophysiological (fMRI) time series analysis: resampling methods in time and wavelet domains., *Human Brain Mapping* 12 (2) (2001) 61–78.
- [89] M. J. Hove, M. T. Fairhurst, S. A. Kotz, P. E. Keller, Synchronizing with auditory and visual rhythms: An fMRI assessment of modality differences and modality appropriateness., *Neuroimage* 67 (2013) 313–321.
- [90] Neurobehavioral systems presentation software Software available at <http://www.neurobs.com>.
- [91] J. P. Marques, T. Kober, G. Krueger, W. van der Zwaag, P.-F. Van de Moortele, R. Gruetter, MP2RAGE, a self bias-field corrected sequence for improved segmentation and T1-mapping at high field, *Current opinion in neurobiology* 49 (2) (2010) 1271–1281.
- [92] R. M. Heidemann, A. Anwender, T. Feiweier, T. R. Knösche, R. Turner, k-space and q-space: Combining ultra-high spatial and angular resolution in diffusion imaging using ZOOPPA at 7T, *Current opinion in neurobiology* 60 (2) (2012) 967–978.

- [93] K. J. Friston, S. Williams, R. Howard, R. Frackowiak, R. Turner, Movement-related effects in fMRI time-series, *Magnetic resonance in medicine* 35 (3) (1996) 346–355.
- [94] M. Bianciardi, M. Fukunaga, P. van Gelderen, S. G. Horovitz, J. A. de Zwart, K. Shmueli, J. H. Duyn, Sources of functional magnetic resonance imaging signal fluctuations in the human brain at rest: a 7 T study, *Magnetic Resonance Imaging* 27 (8) (2009) 1019–1029.
- [95] M. D. Fox, M. E. Raichle, Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging, *Nature Reviews Neuroscience* 8 (9) (2007) 700–711.
- [96] W. Chau, A. R. McIntosh, The Talairach coordinate of a point in the MNI space: how to interpret it., *Neuroimage* 25 (2) (2005) 408–416.
- [97] A. Nieto-Castañón, S. S. Ghosh, J. A. Tourville, F. H. Guenther, Region of interest based analysis of functional imaging data., *Neuroimage* 19 (4) (2003) 1303–1316.
- [98] M. Brett, I. S. Johnsrude, A. M. Owen, The problem of functional localization in the human brain., *Nature Reviews Neuroscience* 3 (3) (2002) 243–249.
- [99] N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 2000.
- [100] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology* 2 (2011) 27:1–27:27, software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [101] C.-C. Chang, C.-J. Lin, LIBSVM, *ACM Transactions on Intelligent Systems and Technology* 2 (3) (2011) 1–27.
- [102] S. Bode, J.-D. Haynes, Decoding sequential stages of task preparation in the human brain., *Neuroimage* 45 (2) (2009) 606–613.
- [103] H. Abdi, L. J. Williams, *Principal component analysis*, Wiley Interdisciplinary Reviews: Computational Statistics 2 (4) (2010) 433–459.
- [104] J. D. Carlin, J. B. Rowe, N. Kriegeskorte, R. Thompson, A. J. Calder, Direction-sensitive codes for observed head turns in human superior temporal sulcus., *Cerebral cortex (New York, NY : 1991)* 22 (4) (2012) 735–744.
- [105] T. Kahnt, J. Heinzle, S. Q. Park, J.-D. Haynes, The neural code of reward anticipation in human orbitofrontal cortex, *Proceedings of the National Academy of Sciences of the United States of America* 107 (13) (2010) 6010–6015.
- [106] W. Sato, T. Kochiyama, S. Uono, S. Yoshikawa, Commonalities in the neural mechanisms underlying automatic attentional shifts by gaze, gestures, and symbols, *Neuroimage* 45 (3) (2009) 984–992.

- [107] P. P. Thakral, S. D. Slotnick, The role of parietal cortex during sustained visual spatial attention, *Brain research* 1302 (C) (2009) 157–166.
- [108] F. Pereira, T. Mitchell, M. Botvinick, Machine learning classifiers and fMRI: A tutorial overview, *Neuroimage* 45 (1) (2009) S199–S209.
- [109] E. Zarahn, G. K. Aguirre, M. D’Esposito, Empirical analyses of BOLD fMRI statistics. I. Spatially unsmoothed data collected under null-hypothesis conditions, *Neuroimage* 5 (3) (1997) 179–197.
- [110] A. Isaksson, M. Wallman, H. Goransson, M. Gustafsson, Cross-validation and bootstrapping are unreliable in small sample classification, *Pattern Recognition Letters* 29 (14) (2008) 1960–1965.
- [111] U. Wickenberg-Bolin, H. Göransson, M. Fryknäs, M. G. Gustafsson, A. Isaksson, Improved variance estimation of classification performance via reduction of bias caused by small sample size, *BMC bioinformatics* 7 (2006) 127.
- [112] A. Hefny, A New Monte Carlo-Based Error Rate Estimator, *Artificial Neural Networks in Pattern Recognition* 1 (2010) 1.
- [113] O. Ibe, *Fundamentals of Applied Probability and Random Processes*, Elsevier Academic Press, 2005.
- [114] M. Hisakado, K. Kitsukawa, S. Mori, Correlated binomial models and correlation structures, *Journal of Physics A: Mathematical and General* 39 (2006) 15365.
- [115] B. Efron, R. Tibshirani, J. D. Storey, V. Tusher, Empirical Bayes Analysis of a Microarray Experiment, *Journal of the American Statistical Association* 96 (456) (2001) 1151–1160.
- [116] K. Strimmer, *fdrtool: a versatile R package for estimating local and tail area-based false discovery rates*, *Bioinformatics* 24 (12) (2008) 1461–1462.
- [117] P. Golland, B. Fischl, Permutation tests for classification: towards statistical significance in image-based studies., *Information processing in medical imaging* 18 (2003) 330–341.
- [118] R. Heller, D. Stanley, D. Yekutieli, N. Rubin, Cluster-based analysis of FMRI data, *Neuroimage* 33 (2006) 599–608.
- [119] S. D. Forman, J. D. Cohen, M. Fitzgerald, W. F. Eddy, M. A. Mintun, D. C. Noll, Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold., *Magnetic resonance in medicine* 33 (5) (1995) 636–647.
- [120] S. M. Smith, T. E. Nichols, Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference., *Neuroimage* 44 (1) (2009) 83–98.

- [121] B. Thirion, P. Pinel, S. Mériaux, A. Roche, S. Dehaene, J.-B. Poline, Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses, *Neuroimage* 35 (1) (2007) 105–120.
- [122] J. Stelzer, T. Buschmann, G. Lohmann, D. S. Margulies, R. Trampel, R. Turner, Prioritizing spatial accuracy in high-resolution fMRI data using multivariate feature weight mapping, *Frontiers in Neuroscience (Brain Imaging Methods)* 8.
- [123] S. Viswanathan, M. Cieslak, S. T. Grafton, On the geometric structure of fMRI searchlight-based information maps, *arXiv.org* 1 (1) (2012) 1.
- [124] S. B. Laughlin, Communication in Neuronal Networks, *science* 301 (5641) (2003) 1870–1874.
- [125] J. Diedrichsen, T. Wiestler, J. W. Krakauer, Two distinct ipsilateral cortical representations for individuated finger movements., *Cerebral cortex (New York, NY : 1991)* 23 (6) (2013) 1362–1377.
- [126] J. A. Etzel, J. M. Zacks, T. S. Braver, Searchlight analysis: Promise, pitfalls, and potential, *Neuroimage* (2013) 1–9.
- [127] J. D. Horn, S. T. Grafton, M. B. Miller, Individual Variability in Brain Activity: A Nuisance or an Opportunity?, *Brain Imaging and Behavior* 2 (4) (2008) 327–334.
- [128] M. B. Miller, J. D. Van Horn, G. L. Wolford, T. C. Handy, M. Valsangkar-Smyth, S. Inati, S. Grafton, M. S. Gazzaniga, Extensive Individual Differences in Brain Activations Associated with Episodic Retrieval are Reliable Over Time, *Journal of Cognitive Neuroscience* 8 (2002) 1200–1214.
- [129] M. B. Miller, C.-L. Donovan, J. D. Van Horn, E. German, P. Sokol-Hessner, G. L. Wolford, Unique and persistent individual patterns of brain activity across different memory retrieval tasks., *Neuroimage* 48 (3) (2009) 625–635.
- [130] M. B. Miller, C.-L. Donovan, C. M. Bennett, E. M. Aminoff, R. E. Mayer, Individual differences in cognitive style and strategy predict similarities in the patterns of brain activity between individuals, *Neuroimage* 59 (1) (2012) 83–93.
- [131] H. Liu, Z. Liu, M. Liang, Y. Hao, L. Tan, F. Kuang, Y. Yi, L. Xu, T. Jiang, Decreased regional homogeneity in schizophrenia: a resting state functional magnetic resonance imaging study, *Neuroreport* 17 (1) (2006) 19.
- [132] J. Gonzalez-Castillo, Z. S. Saad, D. A. Handwerker, S. J. Inati, N. Brenowitz, P. A. Bandettini, Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis., *Proceedings of the National Academy of Sciences of the United States of America* 109 (14) (2012) 5487–5492.